A skeptical appraisal of the bootstrap approach in fund performance evaluation

Huazhu Zhang, Cheng Yan*

8 September 2017

Abstract

It has become standard practice in the fund performance evaluation literature to use the bootstrap approach to distinguish "skills" from "luck", while its reliability has not been subject to rigorous statistical analysis. This paper reviews and critiques the bootstrap schemes used in the literature, and provides a simulation analysis of the validity and reliability of the bootstrap approach by applying it to evaluating the performance of hypothetical funds under various assumptions. We argue that this approach can be misleading, regardless of using alpha estimates or their t-statistics. While alternative bootstrap schemes can result in improvements, they are not foolproof either. The case can be worse if the benchmark model is misspecified. It is therefore only with caution that we can use the bootstrap approach to evaluate the performance of funds and we offer some suggestions for improving it.

Keywords: Monte Carlo simulation; Performance evaluation; Bootstrapping; Fama–French model; Alpha.

JEL classification: C15; G11

---

* Corresponding author: Cheng Yan, Lecturer in Finance, Durham University Business School, Millhill Lane, Durham, UK, DH1 3LB; +44 (0) 191 3345 197; E-mail: cheng.yan@durham.ac.uk. Huazhu Zhang, Nomura. The authors have contributed equally to this work. We are grateful to Jushan Bai, David Blake, Chris Brooks, Biqing Cai, Yong Chen, Tingting Cheng, Ken French, Jiti Gao, Stewart Hodges, Robert Kosowski, Yan Liu, Alex Stremme, Lubos Pastor, and others in various conferences and workshops for helpful comments and discussions. The views expressed herein are those of the authors and do not necessarily reflect or represent those of Nomura.

"An extensive literature in financial economics has focused on the question of whether stock picking or market timing talent exists. Interestingly, the literature has not been able to provide a definitive answer to this question." (Berk and van Binsbergen, 2015)

1. Introduction

Fund performance has been under intensive scrutiny from both practitioners and researchers (for a relatively recent survey, see, e.g., Ferson, 2010; or Wermers, 2011). Acknowledging the weakness of standard parametric t-tests (e.g., Jensen, 1968; Malkiel, 1995) and persistence tests (e.g., Carhart, 1997), a strand of literature stemming from Kosowski et al. (2006) applies the bootstrap technique to identify "skills" (abnormal returns, or "*alpha*") among a large sample of funds and report findings that are strikingly different from those documented in the classical literature. Unlike the standard t-test, this approach requires no *ex ante* parametric assumptions on fund alphas. It allows for the generation of the cross-sectional distribution of fund alphas purely due to sampling variability ("luck"), against which, the cross-section of realized alphas obtained from estimating a linear factor model is compared. Significant difference between them is regarded as evidence of genuine skills. So far, this approach has been extensively used in distinguishing skills from luck in mutual funds (e.g., Cuthbertson et al., 2008; Fama and French, 2010; Blake et al., 2014, 2017), Hedge funds (e.g., Kosowski et al., 2007), and pension funds (e.g., Blake et al., 2013).

Given the popularity and the seeming superiority of the bootstrap approach in distinguishing skills from luck, it is surprising that little rigorous statistical analysis has been conducted to examine whether this approach can actually lead to correct inferences on fund skills as presumed (Cheng and Yan, 2017). Such analysis is of essential importance to the understanding of the recent literature and the appropriate application of bootstrap in future research. We fill this gap. Using Monte Carlo simulations, we evaluate its performance in detecting skills across a large number of hypothetical funds under varieties of assumptions on fund alphas and errors in the linear factor models.

Our research is also motivated by the mixed empirical results regarding mutual funds. We were puzzled by the fact that Kosowski et al. (2006) and Fama and French (2010) use similar bootstrap specifications but reach diverging conclusions, both of which contradict Berk and Green (2004) and Barras et al. (2010). On the one hand, while Kosowski et al. (2006) show that the top decile US mutual funds possess superior skills (*positive alphas*) after-cost and the superior skills persist, Fama and French (2010) find that few US mutual funds have the skills to cover their costs (*negative alphas*). On the other hand, Barras et al. (2010) find only 0.6% of US mutual funds are truly skillsed and little performance persistence after controlling for "false discoveries", which is consistent with Berk and Green (2004) who predict that most funds have just sufficient skills (*zero alphas*) to cover their costs.

In our simulations, we mainly focus on the performance of one particular bootstrap scheme which is the independent sampling of fund residuals (Kosowski et al., 2006), under three different Data Generating Processes (DGPs): (i) all funds are endowed with zero alphas (i.e., a scenario of no true fund skills); (ii) all fund alphas are generated randomly from a normal distribution with zero mean (i.e., a scenario of nonzero true fund skills); (iii) 10% of funds have alphas randomly generated from a normal distribution with zero mean while the rest endowed with zero alphas (i.e., a small fraction of funds have nonzero true skills). We generate 273 monthly excess returns for each of 2000 hypothetical funds from various asset pricing models such as the market model (one-factor CAPM) from Jensen (1968), the four-factor model from Carhart (1997). The number of funds and the sample length are chosen to match the previous literature (i.e., Fama and French, 2010).

The statistical inference of bootstrap evaluation approach can be based on the cross-sectional distributions of either alpha estimates or their t-statistics. The latter (i.e., the t-

statistics of alpha estimates) are preferred than the former (i.e., the alpha estimates) as the econometrics literature suggests that the t-statistics correct for idiosyncratic risk and are more stable cross-sectionally distributed (e.g., Kosowski et al., 2006; Fama and French, 2010). We provide a comparison of them, as it is not clear, at least to us, whether these properties automatically translate into advantages of the t-statistics in distinguishing skills from luck.

We also examine the performance of three alternative bootstrap schemes: a joint sampling of fund residuals, a joint sampling of factor returns and fund residuals, as well as block sampling of fund residuals and factor returns. The alternative bootstrap schemes can either capture correlations across fund returns/residuals or preserve the dependence structure between the factor returns and residuals. While Kosowski et al. (2006) jointly bootstrap fund residuals, Fama and French (2010) jointly bootstrap factor returns and fund residuals. We are interested in knowing whether these schemes can make differences relative to the independent sampling of fund residuals and reconcile the mixed findings from Kosowski et al. (2006) and Fama and French (2010).

Several other concerns also necessitate a statistical analysis of the reliability of the bootstrap evaluation approach. First, it uses OLS regression to estimate fund alphas and betas, which suffer from estimation errors. Second, violations of regression assumptions may cause biases or inefficiency in coefficient estimates and compromise the reliability of this approach. Third, the limited length of fund returns may further aggravate the issue of estimation risk. Finally, benchmark misspecification may also cause problems.

We report four main conclusions. First of all, the bootstrap evaluation approach may falsely identify fund skills and thus cast doubts on its previous applications. On the one hand, even when all funds are endowed with zero alphas in our Data Generating Processes (DGPs), the bootstrap evaluation approach may indicate superior skills among top funds. On the other

hand, if all funds are endowed with well-defined non-zero alphas, it can effectively establish the existence of skills, superior or bad. However, if the distribution of the errors terms exhibits much higher dispersion than the one of the alphas in DGPs, the bootstrap evaluation approach may also fail to detect the existence of skills.

Second, the use of the t-statistics of alpha estimates rather than the alpha estimates themselves in the bootstrap evaluation approach does not guarantee any advantage in terms of validity and reliability. On the contrary, due to the strong robustness property of the t-statistics of alpha estimates, a focus on t-statistics alone may hide the potential issues that can be detected by examining alpha estimates sometimes.

Third, alternative bootstrap schemes that correct for correlations and preserve data dependence structure tend to shrink the cross-sectional distribution of bootstrapped alphas and can result in improvements in statistical inferences on fund skills. Nevertheless, these alternative bootstrap schemes are not foolproof either.

Finally, we also investigate the possible consequences brought by the violations of classical OLS regression assumptions in the bootstrap evaluation framework, as the bootstrap evaluation approach is partially motivated by the fact that the classical OLS regression assumptions are often violated in earlier fund evaluation studies. We find that, in the presence of most problems (i.e., serial correlation, GARCH effect and contemporaneous correlation among errors; and multicollinearity in factor returns), the bootstrap evaluation approach may falsely identify fund skills when all funds are endowed with zero alphas. The consequence is especially severe in the presence of some omitted factor, as the bootstrap evaluation approach may mistake superior (inferior) skills for inferior (superior) due to underestimating (overestimating) realized alphas via OLS.

Overall we find that the bootstrap evaluation approach hinges critically on the appropriateness of using likelihoods as the p-values for statistical inferences on fund skills. Theoretically, we could obtain the empirical distribution of likelihoods at each percentile by generating a large number of paths of fund returns, and compare the likelihoods obtained from historical returns against its empirical distribution to test for the difference between realized and bootstrapped alphas. However, this would be computationally intensive and unnecessary. Once the distribution of the realized alphas is obtained, a simple variance decompositions and Kolmogorov-Smirnov tests can come to rescue when the bootstrap evaluation approach fails. When all funds are endowed with zero alphas, even if the bootstrap evaluation approach tends to predict skills, the variance ratio will be very close to one and the Kolmogorov-Smirnov test will not reject the null of no difference between the distributions of bootstrapped and realized alphas.

To the best of our knowledge, we are the first to doubt the validity and reliability of the bootstrap approach for distinguishing skills from luck in fund performance evaluation from the perspective of "likelihoods". Fama and French (2010) examine how much alpha is necessary to reproduce the cross-section of t-statistics of alpha estimates for actual gross fund returns, but do not address our questions mentioned above. They simulate an alpha for each fund from a normal distribution with a zero mean and standard deviation ranging from 0% to 2% in steps of 0.5% per year and compare the percentiles of so generated alphas with those of realized alphas. While Fama and French (2010) show the level of standard deviation required to generate percentiles of realized alphas, they remain silent on our question whether the bootstrap evaluation procedure can effectively distinguish skills from luck.

The rest of this paper is organized as follows. Section 2 introduces the bootstrap evaluation approach. Section 3 describes our baseline simulation settings with three DGP

scenarios. Section 4 presents our simulation results for the market model. In Section 5, we present simulation results from the variance decomposition analysis, Kolmogorov-Smirnov tests, and the bootstrap evaluation for the four-factor model. Section 6 investigates the possible consequences brought by the violations of classical OLS regression assumptions. Section 7 shows results from the bootstrap evaluation under three alternative bootstrap schemes when the four-factor model is used as the benchmark. Section 8 concludes**.**

2. The Bootstrap Evaluation Approach

In this section, we introduce the bootstrap approach. We start with a description of the benchmark linear factor models for estimating fund alphas, and provide a formal analysis of realized alphas, error-induced alphas, and their relationship as one of our contributions to improving our understanding of the bootstrap evaluation approach. We detail the bootstrap evaluation procedures and the rationale for the use of "likelihoods" in statistical inferences on fund skills by Kosowski et al. (2006) and Fama and French (2010), among others. We also briefly introduce another two statistical tools as the competitors of the bootstrap approach, variance decomposition and the Kolmogorov-Smirnov test at the end of this section.

2.1 DGP and Benchmark Models for Fund Performance Evaluation

The starting point of almost all fund evaluation approaches is to estimate abnormal returns (i.e., the alphas). There are plenty of linear factors models can be used to achieve this goal, and most papers find that their results are robust to the choice of the benchmark linear factor models. For instance, both Kosowski et al. (2006) and Cuthbertson et al. (2008) report that their findings based on the four-factor model are robust to the conditional beta model (Ferson and Schadt, 1996), the conditional alpha-beta model (Christopherson et al., 1998) and the market-timing models (Treynor and Mazuy, 1966; Henriksson and Merton, 1981).

We conjecture that the choice of the benchmark model may affect the performance of the bootstrap evaluation approach, as the estimation errors deteriorate when the number of parameters to estimate increases. The situation will be particularly serious when the benchmark model is not the DGP model, i.e., the benchmark model is misspecified. Consequently, we have tried every model above but mainly present results from two benchmark models which are also our DGP models to illustrate our ideas: The first one is the market model from Jensen (1968), which is used to generate fund returns and as the performance benchmark in Section 4:

$$r_{it} = \alpha_i + \beta_i r_{mt} + \varepsilon_{it} \tag{1}$$

where r_{it} is the excess return on the ith fund, r_{mt} is the excess return on the market factor, α_i measures the abnormal return, β_i is the factor loading for fund i and errors ε_{it} are $i.i.d.$ with zero mean and variance σ_ε^2.

The second benchmark and DGP model is the commonly used four-factor model from Carhart (1997), which is used to generate fund returns and as the performance benchmark in Section 5:

$$r_{it} = \alpha_i + \beta_{1i} r_{mt} + \beta_{2i} SMB_t + \beta_{3i} HML_t + \beta_{4i} MOM_t + \varepsilon_{it} \tag{2}$$

where SMB_t, HML_t and MOM_t are the size, value and momentum factor, respectively.

To investigate the consequence of an omitted factor from the benchmark model, we present the results using Ferson and Schadt's (1996) conditional beta model to generate fund returns, but the market model (1) as the benchmark fund performance evaluation model in and only in subsection 6.5.

2.2 Realized and Error-Induced Alphas

In this subsection, we provide a formal analysis of realized alphas, error-induced alphas and their implications for skills identification, which is one of our contributions to improving the understanding of the bootstrap evaluation approach.

Realized Alphas

The realized alpha for the ith fund is the regression estimate of the intercept of the benchmark linear factor models, e.g., the market model (1) for brevity:

$$\hat{\alpha}_i = \bar{r}_i - \hat{\beta}_i \bar{r}_m \tag{3}$$

where

$$\hat{\beta}_i = \frac{\sum_{t=1}^{T} r_{it} r_{mt} - T\bar{r}_i \bar{r}_m}{\sum_{t=1}^{T} r_{mt}^2 - T\bar{r}_m^2} \tag{4}$$

$\bar{r}_i = \sum_{t=1}^{T} r_{it} /T$ is the average excess return on the ith fund and $\bar{r}_m = \sum_{t=1}^{T} r_{mt} /T$ is the average excess return on the market factor.

Under classical regression assumptions, $\hat{\alpha}_i$ is an unbiased estimator of α_i and its standard deviation is

$$Std(\hat{\alpha}_i) = \hat{\sigma}_\varepsilon \sqrt{1/T + \bar{r}_m / \sum_{t=1}^{T} (r_{mt} - \bar{r}_m)^2} \tag{5}$$

where

$$\hat{\sigma}_\varepsilon = \left(r_{it} - \hat{\alpha}_i - \hat{\beta}_i r_{mt}\right)^2 / (T-2) \tag{6}$$

The t-statistics of $\hat{\alpha}_i$ is $t_{\hat{\alpha}_i} = \hat{\alpha}_i / Std(\hat{\alpha}_i)$. Percentiles of realized alphas and their t-statistics are obtained by ranking $\{\hat{\alpha}_i, i = 1, 2, \dots, N\}$ and $\{t_{\hat{\alpha}_i}, i = 1, 2, \dots, N\}$.

Error-Induced Alphas

The realized alpha is subject to the influence of sampling errors ("luck"). To isolate the effect of sampling errors ("luck"), we remove the actual α_i from fund returns r_{it}

$$r_{it}^E = r_{it} - \alpha_i = \beta_i r_{mt} + \varepsilon_{it} \tag{7}$$

and then re-estimate the market model using r_{it}^E in place of r_{it} in (1). We obtain

$$\hat{\alpha}_i^E = \bar{r}_i^E - \hat{\beta}_i^E \bar{r}_m = (\beta_i - \hat{\beta}_i)\bar{r}_m + \bar{\varepsilon}_i \tag{8}$$

where

$$\hat{\beta}_i^E = \frac{\sum_{t=1}^T r_{it}^E r_{mt} - T\bar{r}_i^E \bar{r}_m}{\sum_{t=1}^T r_{mt}^2 - T\bar{r}_m^2} = \hat{\beta}_i \tag{9}$$

$\bar{r}_i^E$ and $\bar{\varepsilon}_i$ are respectively the averages of r_{it}^E and ε_{it}.[1]

We call $\hat{\alpha}_i^E$ the error-induced alpha since it is purely due to sampling errors ("luck"). Its mean and variance are zero and $Var(\hat{\beta}_i)\bar{r}_m^2 + \sigma_\varepsilon^2/T$, respectively.[2] It is related to the realized alpha in the following way[3]

$$\hat{\alpha}_i = \hat{\alpha}_i^E + \alpha_i \tag{10}$$

The above equation has two important implications. First, the distribution of realized alphas $\{\hat{\alpha}_i, i = 1, 2, \dots, N\}$ is the superimposition of those of error-induced alphas $\{\hat{\alpha}_i^E, i = 1, 2, \dots, N\}$ ("luck") and true alphas $\{\alpha_i, i = 1, 2, \dots, N\}$ ("skills"). If $\alpha_i = \alpha, \forall i$, that is, all funds have identical true alpha α ("skills"), the distribution of realized alphas will be a translation of the error-induced one by a constant α. In case $\alpha_i = 0, \forall i$, these two distributions coincide. Second, the variance of $\hat{\alpha}_i$ partitions into the variance of $\hat{\alpha}_i^E$ and that of α_i. When true alphas are zero, the variation of $\hat{\alpha}_i$ can be fully explained by the variation of

---

[1] Proofs of (9) and (8): $\hat{\beta}_i^E = \frac{\sum_{t=1}^T r_{it}^E r_{mt} - T\bar{r}_i^E \bar{r}_m}{\sum_{t=1}^T r_{mt}^2 - T\bar{r}_m^2} = \frac{\sum_{t=1}^T (r_{it} - \alpha_i) r_{mt} - T(\bar{r}_i - \alpha_i)\bar{r}_m}{\sum_{t=1}^T r_{mt}^2 - T\bar{r}_m^2} = \frac{\sum_{t=1}^T r_{it} r_{mt} - T\bar{r}_i \bar{r}_m}{\sum_{t=1}^T r_{mt}^2 - T\bar{r}_m^2} = \hat{\beta}_i$;

$\hat{\alpha}_i^E = \bar{r}_i^E - \hat{\beta}_i^E \bar{r}_m = \beta_i \bar{r}_m + \bar{\varepsilon}_i - \hat{\beta}_i \bar{r}_m = (\beta_i - \hat{\beta}_i)\bar{r}_m + \bar{\varepsilon}_i$.

[2] $E[\hat{\alpha}_i^E] = E[\hat{\alpha}_i - \alpha_i] = E[\hat{\alpha}_i] - \alpha_i = 0$; $Var(\hat{\alpha}_i^E) = Var[(\beta_i - \hat{\beta}_i)\bar{r}_m] + Var[\bar{\varepsilon}_i] = Var(\hat{\beta}_i)\bar{r}_m^2 + \sigma_\varepsilon^2/T$.

[3] Proof of (10): $\hat{\alpha}_i = \bar{r}_i - \hat{\beta}_i \bar{r}_m = \bar{r}_i^E + \alpha_i - \hat{\beta}_i^E \bar{r}_m = \hat{\alpha}_i^E + \alpha_i$.

$\hat{\alpha}_i^E$. If we take the ratio of the two variances, we should expect a value very close to one. However, a ratio close to one may also arise from other possibilities. For example, if $\hat{\alpha}_i^E$ has a very large dispersion relative to α_i, the dispersion of $\hat{\alpha}_i$ will be largely determined by that of $\hat{\alpha}_i^E$, in which case, the magnitudes of alpha estimates will become important in detecting fund skills.

2.3 Bootstrapped Alphas

We see from (8) that the computation of the error-induced alpha $\hat{\alpha}_i^E$ requires the knowledge of β_i and $\bar{\varepsilon}_i$, neither of which is observable. We can use bootstrapping to estimate error-induced alphas and call them bootstrapped alphas.

To avoid possible ambiguities, we write down explicitly the steps that Kosowski et al. (2006) use to obtain bootstrapped alphas.[4] We first estimate the market model (1) for each fund and save the vector of estimates $\{\hat{\alpha}_i, \hat{\beta}_i, \hat{\varepsilon}_{it}, t_{\hat{\alpha}_i}\}$, in which, $\hat{\alpha}_i$ is the realized alpha, $t_{\hat{\alpha}_i}$ is the t-statistic of $\hat{\alpha}_i$, $\hat{\beta}_i$ is the estimate of the factor loading and $\hat{\varepsilon}_{it}$ is the fund residual. Then we follow the procedures stated below to obtain bootstrapped alphas

1. draw randomly a sample of length T with replacement from fund residuals $\{\hat{\varepsilon}_{it}, t = 1, 2, \dots, T\}$ and denote them as $\hat{\varepsilon}_{it}^b$;
2. obtain bootstrapped returns $r_{it}^b = \hat{\beta}_i r_{mt} + \hat{\varepsilon}_{it}^b$ under the null hypothesis $\alpha_i = 0$;
3. use $\{r_{it}^b, t = 1, \dots, T\}$ to re-estimate the market model (1) and denote the resulting alpha estimate and its t-statistics by $\hat{\alpha}_i^b$ and $t_{\hat{\alpha}_i}^b$, respectively;

[4] Similar procedures are used by Cuthbertson et al. (2008) and Fama and French (2010). There is a slight difference Kosowski et al. (2006) and Fama and French (2010), as Kosowski et al. (2006) bootstrap fund residuals and factor returns independently while Fama and French (2010) sample them jointly.

4. rank $\{\hat{\alpha}_i^b, i = 1, 2, \dots, N\}$ and $\{t_{\hat{\alpha}_i}^b, i = 1, 2, \dots, N\}$ to obtain their respective percentiles $\{\hat{\alpha}_p^b, p = 0.01, 0.02, \dots, 0.99\}$ and $\{t_p^b, p = 0.01, 0.02, \dots, 0.99\}$.

We repeat the above procedures $B(= 5000)$ times and compute the percentiles for bootstrapped alphas and their t-statistics as follows

$$\bar{\alpha}_p^B = \frac{1}{B}\sum_{b=1}^{B} \hat{\alpha}_p^b \, , p = 0.01, 0.02, \dots, 0.99 \tag{11}$$

$$\bar{t}_p^B = \frac{1}{B}\sum_{b=1}^{B} t_p^b \, , p = 0.01, 0.02, \dots, 0.99 \tag{12}$$

Let $\hat{\alpha}_p^R$ and t_p^R denote the pth percentile of realized alphas and their t-statistics. We obtain them by ranking realized alphas $\{\hat{\alpha}_i, i = 1, 2, \dots, N\}$ and their t-statistics $\{t_{\hat{\alpha}_i}, i = 1, 2, \dots, N\}$. Let $\mathcal{L}_p$ denote the proportion of bootstrapped alphas out of B bootstraps that fall below the realized alpha at the pth percentile. We have

$$\mathcal{L}_p = \frac{\sum_{b=1}^{B} \mathbf{1}(\hat{\alpha}_p^b < \hat{\alpha}_p^R)}{B}, p = 0.01, 0.02, \dots, 0.99 \tag{13}$$

where $\mathbf{1}(\hat{\alpha}_p^b < \hat{\alpha}_p^R)$ takes value one if $\hat{\alpha}_p^b < \hat{\alpha}_p^R$ and zero otherwise.

Similarly, we can obtain the proportion of t-statistics of bootstrapped alphas out of B bootstraps that fall below that of the realized alpha at the pth percentile.

2.4 Percentiles, Likelihoods, and Potential Problems

To separate skills from luck, the current literature focuses on comparing the values of realized and bootstrapped alphas at selected percentiles with particular emphasis on the extreme tails. $\mathcal{L}_p$ in (13) is called the "likelihood" by Fama and French (2010) and plays a pivotal role in their statistical inference on skills.

The rationale for likelihood-based inferences is as follows. Bootstrapped alphas are purely due to sampling variations and thus represent luck while realized alphas are driven by both luck and true alphas that represent skills. If funds do have the superior good (bad) skills, that is, true alphas are significantly different from zero, we should expect there is a

significant difference between the cross-sections of bootstrapped and realized alphas at the right (left) tail. In the literature, researchers focus on the difference between realized and bootstrapped alphas at selected percentiles, in particular, the top and bottom percentiles. For example, if the likelihood $\mathcal{L}_p$ at the 99th percentile is 95%, which means that 95% of bootstrapped alphas out of the B bootstraps fall below their realized counterpart, then this will be accepted as evidence of superior skills in the literature. Similarly, at bottom percentiles, if the likelihood is significantly small, then this will be regarded as evidence of bad skills.

This likelihood-based inference has at least two problems. First, we use bootstrapped alphas to approximate error-induced alphas. In theory, we need to compare the cross-sectional distributions of realized alphas and error-induced alphas. If all funds just match their benchmark, then the two distributions will coincide; otherwise, the distribution of realized alphas should have longer tails than that of error-induced alphas. In a finite sample, especially when there is an insufficient length of historical data, the distribution of bootstrapped alphas may significantly deviate from that of error-induced alphas, which may cause problems when we make inferences on skills by comparing the cross-section of realized alphas against that of bootstrapped alphas rather than error-induced alphas. As the number of parameters to estimate increases, estimation risk will aggravate, and thus the inference may become even less reliable. The quality of the inference will depend on the position of the distribution of bootstrapped alphas relative to those of error-induced and realized alphas, which can only be analyzed using Monte Carlo simulation in the current multivariate context. Second, the results can be shown for many percentiles, and the likelihoods are correlated. Should we look at likelihoods at each percentile? Fama and French (2010) choose to examine all the likelihoods with emphasis on upper and lower percentiles.

Furthermore, how should we interpret likelihoods? If we observe likelihood as good as 95% at the 99th percentile, is it correct to claim that the top 1% of funds have superior skills? Fama and French (2010) appear to have no doubt in this; however, our subsequent analysis shows that this interpretation can lead to problematic inferences on fund skills.

2.5 Inferences Based on Alpha Estimates versus their t-statistics

In the current literature, the likelihood-based inference mostly focuses on t-statistics of fund alphas rather than alpha estimates. Kosowski et al. (2006) argue that t-statistics normalize alpha estimates by their standard deviation, and the cross-sectional distribution of t-statistics is more stable than that of alphas in the presence of heterogeneous fund volatilities. It is well known that under classical regression assumptions, the asymptotic distribution of t-statistics is the standard normal. In contrast, alpha estimates are highly influenced by random errors. As the standard deviation of errors rises, the dispersion of alpha estimates will increase. In contrast, the cross-section of t-statistics is almost immune from this. However, given the complexity of the bootstrap evaluation approach, it is not clear that these nice properties of t-statistics can lead to more reliable results than alpha estimates. To investigate this issue, we provide simulation analyses based on both alpha estimates and their t-statistics.

2.6 Alternatives to bootstrap: Variance Decomposition or Kolmogorov-Smirnov Test

As shown in (10), we can break the variance of the realized alpha into two parts: the variance of true alphas and that of bootstrapped alphas. If the realized alpha is simply due to random errors, the ratio of the variance of bootstrapped alphas over that of realized alphas should be close to one, which means that the well-known variance decomposition can help identify fund skills. Surprisingly, the literature focuses on the first moment of bootstrapped and realized alphas, but make no direct use of variances of bootstrapped and realized alphas. We, therefore, show how such information can be used to shed light on identifying fund skills.

Ideally, we should make formal statistical inferences on skills based on the distribution of the variance decomposition ratio. Unfortunately, its probability distribution is beyond our knowledge. Therefore, we just report the ratios obtained from kernel density estimates of the variances of the distributions of realized and bootstrapped alphas.

A well-known formal statistical test for difference between two distributions is the Kolmogorov-Smirnov two-sample test. It is based on the maximal difference between cumulative frequency distributions of two samples and does not require the specification of the underlying distribution. We use it to test for difference between the cross-sectional distributions of realized alphas and bootstrapped alphas.

3. Simulation Settings

We generate N=2000 hypothetical funds and use B=5000 bootstraps to compute bootstrapped alphas. For each fund, we generate T=273 monthly returns from either the market model (1) or the four-factor model (2).[5] In the following, we introduce three DGP scenarios for the composition of fund alphas, how to generate monthly returns using the market model and the four-factor model, three alternative bootstrap schemes under our consideration and alternative probability distributions that we use to simulate errors in the benchmark models.

[5] The maximum number of funds covered by Kosowski et al. (2006) is 1788, while Fama and French (2010) include 3,156 funds in the $5 million group, 1,422 in the $250 million group, and 660 in the $1 billion group. The maximum length of past returns that an individual fund can have is 336 and 273 in Kosowski et al. (2006) and Fama and French (2010), respectively. We take the sample of 273 months from Fama and French (2010) but check a sample length up to 2000 whenever possible.

3.1 Three DGP scenarios for Fund Alphas

We analyze the performance of the bootstrap evaluation approach under three different Data Generating Process (DGP) scenarios for true alphas of the 2000 hypothetical funds.

DGP 1: $\boldsymbol{\alpha_i = 0, \forall i}$

This case is similar to Berk and Green (2004). In this DGP, all funds are endowed zero alphas. Since true alphas are zero, both realized and bootstrapped alphas are due to sampling variations. Therefore, there should be no significant difference between cross-sectional distributions of realized and bootstrapped alphas and thus their values at selected percentiles. If the bootstrap evaluation approach is valid and reliable, we should expect likelihoods to be nowhere close to zero or one at the selected percentiles. Otherwise, we can conclude that it is not appropriate to use likelihoods in statistical inferences on skills.

DGP 2: $\boldsymbol{\alpha_i \sim N(0, 0.02^2), \forall i}$

This case is similar to Kosowski et al. (2006) and Fama and French (2010). Fund alphas are randomly drawn from the normal distribution with zero mean and annual standard deviation of 2%.[6] Therefore, there exists superior skills among top-performing funds and bad skills among bottom-performing funds. If the likelihood-based inference is valid and reliable, we will see a sharp difference between realized alphas and their bootstrapped values and likelihoods very close to one at upper percentiles and virtually zero at lower percentiles.

DGP 3: $\boldsymbol{\alpha_i \sim N(0, 0.02^2), i = 1, \ldots, 200; \alpha_i = 0, i = 201, \ldots, 2000}$

[6] Fama and French (2010) examine how much alpha is necessary to reproduce the cross-section of t-statistics of alpha estimates for actual gross fund returns. They find that the reasonable levels for the standard deviation of true alphas range from 0.5% to 2.0%. We set the standard deviation of true alpha to be the upper level 2.0% to make sure that funds have sufficiently large but also reasonable alphas to allow a normal environment for the bootstrap evaluation method to work in.

This DGP is a mixture of DGP 1 and DGP 2, and similar to Barras et al. (2010). In this DGP, 200 out of 2000 funds have alphas randomly generated from $N(0, 0.02^2)$ and the other funds have zero alphas. For the 200 nonzero alphas, we make half of them positive and the other half negative. Given nonzero alphas for only 10% of the funds, we will see some difference between realized and bootstrapped alphas at top and bottom percentiles if the likelihood-based inference approach is valid and reliable. In terms of likelihoods, they should be close to 100% at top percentiles and virtually zero at bottom percentiles.

3.2 Specifications for Other Parameters in Benchmark Models

Having specified these three DGP scenarios for fund alphas, we detail how to obtain fund betas, factor returns, and errors in the benchmark models in this subsection. Depending on the benchmark model that we use to generate fund returns, there is a slight difference regarding parameter specifications.

In Section 4, the market model (1) is used to generate fund returns and as the performance benchmark. Fund betas, market returns and errors in the market model are simulated as follows

(i). β_i is simulated from a uniform distribution on the interval $[0.5, 2]$

(ii). r_{mt} is simulated from $N(0.08, 0.15^2)$, the normal distribution with mean 0.08 per year and annual standard deviation 15%

(iii). ε_{it} is drawn from $N(0, 0.08^2)$, the normal distribution with zero mean and annual standard deviation 8%

In Section 5, we use the four-factor model (2) to generate fund returns and as the performance benchmark. We obtain data on market excess returns and returns on size, value and momentum factors from French's online data library.[7] The data cover the period January

[7] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

1984 to September 2006, the same as Fama and French (2010). Fund betas with respect to the four factors and errors in the four-factor model are simulated in the same way as described in (i) and (iii).

Given the generated fund returns, we follow the same procedures outlined in section 2.3 to obtain realized and bootstrapped alphas, their percentiles and likelihoods at the selected percentiles.

3.3 Alternative Bootstrap Schemes

At step 2 in Section 2.3, we bootstrap fund residuals independently to obtain bootstrapped returns r_{it}^{b} in the first place. Fama and French (2010) sample fund residuals and factor returns jointly while Kosowski et al. (2006) do simulations independently for each fund and jointly for fund residuals only. In addition to the difference in the sample periods covered by them, Fama and French (2010) argue that their different bootstrap scheme is also part of the reason for the difference between their findings and those of Kosowski et al. (2006).

We also consider three alternative bootstrap schemes: a joint sampling of fund residuals, a joint sampling of fund residuals and factor returns, and block sampling of fund residuals and factor returns (see, Politis and Romano, 1994). While independent sampling of fund residuals disregards any correlation and dependence among errors and factor returns, these three alternative bootstrap schemes can capture respectively the contemporaneous correlation among fund residuals, the contemporaneous correlation among fund residuals and factor returns, and the serial correlations among fund residuals and factor returns. We test these three bootstrap schemes for the four-factor model, in which dependence and correlations are likely to exist among market data on factor returns.

3.4 Alternative Distributions for Errors in the Benchmark Model

The cross-sectional distribution of realized alphas is usually not normal, see Kosowski et al. (2006). We use three alternative distributions for errors in the benchmark model: uniform, t and Normal-Inverse Gaussian (NIG) distributions to generate non-normality in fund returns.[8]

For comparison, parameters of these three distributions are specified so that they have zero mean and annual standard deviation 8%, the same as the normal distribution given in (iii) in Section 3.2. Specifically, the uniform distribution lies on [-0.04, 0.04], the t distribution has a degree of freedom 3 and the NIG distribution has tail heaviness $\lambda = 10$, location parameter $\mu = 0$, asymmetry parameter $\theta = 0$ and scale parameter $\delta = 0.0053$. For further details on the NIG distribution, please refer to the Appendix.

4. Results for the Market Model from Jensen (1968)

In this section, we first examine how realizations of fund betas and errors in the market model (1) from Jensen (1968) affect the percentiles and likelihoods and report the results for these three DGP scenarios of fund alphas detailed in Section 3.1 when the market model (1) also works as the performance benchmark (i.e., the benchmark model is correctly specified). After that, we report results from the variance decomposition analysis, Kolmogorov-Smirnov tests and the bootstrap evaluation for these three DGPs of fund alphas.

[8] Instead of directly simulating fund alphas from uniform, t or NIG distribution, we focus on simulating errors from these distributions for the following reasons. First, it does allow us to inject non-normality into simulated fund returns. Second, the bootstrap evaluation approach involves using OLS regression, which is more sensitive to the error specifications than to the underlying distribution of fund alphas. Third, in untabulated results which are available from the authors upon request, we have investigated and found that different distributions of fund alphas will affect the performance of the bootstrap evaluation approach.

4.1 Effects of Different Realizations of Fund Betas and Errors

In the literature, mutual fund performance is typically evaluated based on the history of fund returns. Since we generate fund returns from the market model (1), we can have different realizations for parameters in the market model (1) and thus different histories of fund returns. We conduct our analysis based on one particular history of fund returns. While this reduces the computational burden substantially, there is a concern about the sensitivity of the results to the history of fund returns. In the first place, we provide a sensitivity analysis of percentiles and likelihoods with respect to realizations and distributions of fund betas and errors in market model (1).

Table 1 presents the annualized realized (*Rea*) and bootstrapped (*Bstrap*) alphas (in percentages) and likelihoods (*%<Rea*) at the selected percentiles (*Pct*) for DGP 1, in which all funds are endowed zero alphas. The likelihood is the proportion of bootstrapped alphas out of 5000 bootstraps that fall below their realized counterparts at the selected percentiles.

In our simulation, fund returns are generated from the market model (1), in which all funds are endowed zero alphas. Panels A, B and C allow us to examine whether the simulation results are sensitive to different distributions of fund betas. They are based on the same realizations of market returns and errors, drawn randomly from $N(0.08, 0.15^2)$ and $N(0, 0.08^2)$, respectively. However, βs in Panels A, B and C are simulated from $U[0.5, 2]$, $U[0.5, 3.5]$ and $N(1.25, 0.43^2)$, respectively. Panels A and D allow us to investigate whether the results are sensitive to different realizations of errors in the market model. Otherwise identical to Panel A, Panel D is based on different realizations of errors from the same distribution $N(0, 0.08^2)$.

We observe that percentiles and likelihoods are not sensitive to the distribution of β. At the selected percentiles (e.g., Percentile 3%, 4%, 5% and 20%), Panels A and B have the

same realized and bootstrapped alphas, and very close likelihoods even though βs are randomly drawn from two different uniform distributions $U[0.5, 2]$ and $U[0.5, 3.5]$, respectively. In Panel C, βs are simulated from the normal distribution with the same mean and variance as the uniform distribution in Panel A. Again, we find virtually no difference in realized and bootstrapped alphas and negligible difference in likelihoods at the selected percentiles (e.g., Percentile 3%, 4%, 5% and 20%).

However, percentiles of realized alphas are very sensitive to the change in the realization of errors. In Panels A and D, errors in the market model have different realizations from the same distribution $N(0, 0.08^2)$. We find that while bootstrapped alphas are approximately the same in Panels A and D at the selected percentiles (e.g., Percentile 3%, 4%, 5% and 20%), the realized values are apparently different. Accordingly, likelihoods exhibit an obvious difference. In untabulated simulation results which are available from the authors upon request, we find that this remains true even if we extend the sample length from T =273 to 2000.

The sensitivity of percentiles and likelihoods to the realization of errors implies that the likelihoods are not reliable for statistical inference on fund skills. The reason is that the likelihood at a selected percentile is arising from a particular path of fund returns and can be regarded as one realization from its underlying distribution. To make any valid statistical inference, we need to know the distribution for likelihoods at the same percentile. While it is challenging if not impossible to derive the distribution for likelihoods at each percentile analytically, we could use Monte Carlo simulations to estimate it. However, this would be computationally insensitive. We also notice that while bootstrapped alphas are approximately symmetric about zero, realized alphas could be slightly skewed. The reason is that bootstrapped alphas at the selected percentiles are the averages across 5000 bootstraps while

realized alphas are simply based on one realization of asset returns generated from the market model and thus suffer more severely from estimation errors.

DGPs 2 and 3 result in similar inferences and are not reported here.

4.2 Results under Classical Regression Assumptions

In the previous subsection, we show that percentiles and likelihoods are very sensitive to the history of generated fund returns. To highlight the potential issues of the bootstrap evaluation approach, in this subsection, we base our analyses on a carefully chosen path of fund returns generated from the market model and present the results for these three DGP scenarios of fund alphas described in Section 3.1 under classical regression assumptions.[9] In addition to the normal distribution $N(0, 0.08^2)$, we also consider three alternatives distributions of errors in the market model: uniform, t and NIG. Their parameters are set so that errors have zero means and annualized standard deviations 0.08.

DGP 1: $\boldsymbol{\alpha_i = 0, \forall i}$

In this DGP, all funds are endowed with zero alphas and thus no fund has skills. If the likelihood-based approach is valid and reliable, we should see that at both upper and lower percentiles, there is no significant difference between realized and bootstrapped alphas. In terms of likelihoods, they should be nowhere close to 100% or zero.

We find that the graphs of probability density functions, the variance decomposition, and the Kolmogorov-Smirnov test reveal that funds have no skills. However, the likelihood-based approach leads to incorrect inferences on funds skills and thus is not reliable.

To be specific, Figure 1 plots the kernel density estimates of probability density functions (PDF) for realized and bootstrapped alphas. As we see, regardless of the

[9] Actually, this carefully chosen path of fund returns is frequently encountered in our simulation.

distribution of errors, there is no significant difference between their PDFs, which is consistent with the assumption of zero alphas.

Table 2 presents the kernel density estimates of mean and variance of bootstrapped and realized alphas and the results of the Kolmogorov-Smirnov tests for the difference between them. First, regardless of the error distribution, bootstrapped and realized alphas have approximately the same mean and variance estimates. The ratio of variances of bootstrapped and realized alphas are very close to one, which implies the variance of bootstrapped alphas accounts for virtually all the variation in realized alphas. Second, the Kolmogorov-Smirnov test fails to reject the null of no difference between the distributions of realized and bootstrapped alphas under all the four error distributions.

However, the percentiles and likelihoods of realized and bootstrapped alphas lead to problematic inferences on fund skills. Table 3 reports realized and bootstrapped alphas and likelihoods at the selected percentiles for DGP 1 under the four error distributions: uniform, normal, t and NIG in Panels A, B, C and D, respectively.

Fama and French (2010) interpret very large likelihoods at top percentiles as evidence of superior skills and very small likelihoods at bottom percentiles as evidence of bad skills. In Panel A, we see that from the 96^{th} to 99^{th} percentiles, realized alphas are above their bootstrapped values with likelihoods above 95%, indicating that more than 95% of alpha estimates out of the 5000 bootstraps fall below their realized counterparts. According to Fama and French (2010), this would be accepted as strong evidence of superior skills among top funds.

With normal errors in Panel B, realized alphas are slightly below their bootstrapped values at upper percentiles and lower percentiles. However, the corresponding likelihoods indicate no significant evidence of skills at 5% level. In Panel C, errors are drawn randomly

from the *t* distribution. We find that starting from the third percentile, realized alphas are smaller than their bootstrapped values. In particular, at the third, fourth, fifth and sixth percentiles, all the likelihoods are below 5%, implying bad skills. For the NIG errors in Panel D, at the 90^{th} (and 95^{th}) percentile, the bootstrapped alpha 2.07 (2.71) is below its realized counterpart 2.20 (2.88) with the likelihood equal to 98.50% (97.10%), indicating the existence of superior skills. At even higher percentiles, bootstrapped alphas stay below their realized counterparts for at least 80% of the 5000 bootstraps.

The tests used by Fama and French (2010), among others, lead to the inference that there is superior and bad skills among funds at extreme tails. The cause for this incorrect inference lies in the misinterpretation of likelihoods. Fama and French (2010) treat likelihoods as if they were the p-values commonly used in statistical tests, which is common in the literature using the bootstrap evaluation approach. However, this interpretation is not appropriate. As we point out in the previous subsection, we need to compare the observed likelihood against its underlying distribution at a particular percentile to make statistical inference on skills. We could use Monte Carlo simulations to estimate the distribution of likelihoods at each percentile and find out the correct p-value of the test; however, this would be computationally intensive. Moreover, the bootstrapped distribution of likelihoods may be very sensitive to the assumptions on parameters and is of little value in empirical performance evaluation.

Table 4 shows the results based on t-statistics of alpha estimates. We find that t-statistics have no advantage over alpha estimates. The patterns of likelihoods based on t-statistics are very similar to what we observe in Table 3. Likelihoods indicate substantial evidence of inferior skills at third, fourth and fifth percentiles in Panel C and evidence of superior skills at the 95^{th} percentile in Panel D.

In summary, when all funds are endowed with zero alphas, the bootstrap evaluation approach can conclude the existence of skills. This is true regardless of using alphas or their t-statistics. In contrast, the graphs of PDFs, the variance decomposition, and the Kolmogorov-Smirnov test result in the correct inference. The cause for the incorrect inferences resulting from the bootstrap evaluation lies in the misinterpretation of likelihoods. Given the history of fund returns, the likelihood obtained at each selected percentile is just one realization from its underlying distribution; however, most literature that uses this approach treats it as if it were the p-value commonly employed in statistical tests.

DGP 2: $\boldsymbol{\alpha_i \sim N(0, 0.02^2), \forall i}$

In this DGP, fund alphas are generated randomly from the normal distribution with zero mean and annual standard deviation of 2%, which means that almost all funds either outperform or underperform the benchmark. If the bootstrap evaluation approach works effectively, we should expect that at upper percentiles, realized alphas are greater than their bootstrapped values with likelihoods close to 100% and at lower percentiles, realized alphas below bootstrapped values with likelihoods close to zero.

Table 5 reports the results based on alpha estimates and their t-statistics under two different error distributions $N(0, 0.08^2)$ and $N(0,1)$ in Panels A and B, respectively.[10] As before, fund returns are generated from the market model (1). Market returns and fund betas are identical to those used in Table 3.

[10] In untabulated results which are available from the authors upon request, variance decompositions and Kolmogrov-Smirnov tests show strong evidence for the difference between the distributions of realized alphas and bootstrapped alphas for Cases 2 and 3.

In Panel A of Table 5, errors in the market model have an annual standard deviation 8%. In Panel B, we specify a substantial annual standard deviation 100% for errors for two reasons. First, we are interested in knowing if a large dispersion in errors will make it difficult to test for difference between realized and bootstrapped alphas when true alphas of funds are non-zero. This is actually the primary cause for the misgivings in the current case if we interpret likelihoods as Fama and French (2010) do. Second, the current literature emphasizes the advantage of using the t-statistics of alpha estimates instead of alpha estimates themselves for the reason that the distribution of t-statistics is standardized and more stable. However, we show that if we follow Fama and French (2010)'s interpretation of likelihoods, this can be a disadvantage when the dispersion of errors is so large that the t-statistics of alpha estimates at selected percentiles become virtually indistinguishable from those of bootstrapped alphas. As we show later, in such situation, it is better to adhere to alpha estimates. We obtain similar findings for uniform, t or NIG errors and do not report them here.

From Panel A of Table 5 we see that even if we treat likelihoods as if they were p-values, we still can arrive at the correct inferences about fund skills regardless of using alpha estimates or their t-statistics here. We see that at lower percentiles realized alphas (and their t-statistics) are far smaller than their bootstrapped values (and their t-statistics). At upper percentiles, the opposite is true. Moreover, likelihoods are zero below the 50^{th} percentile and 100% above the 50^{th} percentile, implying inferior skills at the bottom and superior skills at the top.

The results from Panel A imply that we can treat likelihoods as if they were p-values without making incorrect inferences on skills here. One reasonable explanation is that had we know the distribution of likelihoods at the selected percentile; we would find a substantial

probability for the likelihood to take the value that we see in Panel A at the selected percentile. In fact, as we generate many different histories of fund returns from the same data-generating process as in Panel A, we find that we always obtain the same pattern for likelihoods.

The sharp distinction between realized and bootstrapped alphas at extreme tails is due to the high dispersion in realized alphas compared to that of bootstrapped values. Bootstrapped alphas have a monthly standard deviation 0.14%, in sharp contrast to 0.58%, the standard deviation of true alphas.[11] The dispersion in realized alphas should be even higher because of estimation errors.

The challenge is whether this approach remains capable of separating skills from luck as the dispersion of errors in the market model increases substantially. To investigate this issue, we increase the annual standard deviation of errors to 100% while fixing the annual standard deviation of fund alphas at 2%.

We present the results in Panel B of Table 5. We see that at all the selected percentiles, the gaps between realized and bootstrapped alphas narrow substantially relative to that we have seen in Panel A. First, the dispersion of bootstrapped alphas increases substantially due to the rise in the standard deviation of errors. As a result, the dispersion of realized alphas also grows but to a lesser extent. For example, for realized alphas, the difference between the 99th and 1st percentiles increases to 102.45 in Panel B from 34 in

[11] The monthly standard deviation for bootstrapped alphas is $\sigma_\varepsilon/\sqrt{12T}$, in which, σ_ε denotes the annual standard deviation of errors and T is the sample length. In the current setting, we have $\sigma_\varepsilon = 0.08$ and $T = 273$, which implies a standard deviation 0.14%. Since actual alphas are simulated from $N(0, 0.02^2)$, their monthly standard deviation is $0.02/\sqrt{12}$ =0.58%. As the sample size T increases, the distribution of bootstrapped alphas will shrink further while the distribution of realized alphas, largely determined by that of actual alphas, will not change qualitatively. This means that the separation of skill from luck will become even more obvious.

Panel A. For bootstrapped alphas, the difference is 97.30, compared to 7.79 in Panel A. While at upper percentiles, realized alphas remain above their bootstrapped values, we cannot reject the null of no difference between them at the significance level of 5%.

Inferences based on t-statistics are even worse. First, we see that there is virtually no change in the t-statistics of bootstrapped alphas as we increase the standard deviation of errors from 8% in Panel A to 100% in Panel B. From regression analysis, we know that t-statistics are normalized alpha estimates and follow the standard normal distribution given a large sample, which explains why we see almost no change in the t-statistics of bootstrapped alphas as the standard deviation of errors increases. However, the t-statistics of realized alphas have shrunk substantially. The intuition is that while the standard deviation of realized alphas rises dramatically due to the increase in the standard deviation of errors, realized alphas are determined mainly by true alphas and thus change to a far less extent. As the ratios of realized alphas to their standard deviations, t-statistics will decline dramatically. Actually, as we can see from Panel B, the t-statistics of realized alphas are quite similar to those of bootstrapped alphas, and even closer to those for normal errors in Panel B of Table 4 in which all funds are endowed with zero alphas. As in the case of alpha estimates, the likelihoods imply that we cannot reject the null of no difference between the t-statistics of realized alphas and those of bootstrapped alphas at upper percentiles.

To sum up, given reasonably large errors, the bootstrap evaluation approach works very well when all funds are assumed to either outperform or underperform the market benchmark. Likelihoods can be treated as if they were p-values without causing incorrect inferences on skills here. This is true of inferences based on both alpha estimates and their t-statistics. However, as the standard deviation of errors becomes extremely large, we could draw problematic inferences on skills if we treat likelihoods as if they were p-values for

statistical tests. In this situation, results from alpha estimates are more informative because they reveal the level of dispersion in bootstrapped alphas, which cannot be known from their statistics because their distribution is not sensitive to the change in the standard deviation of errors.

DGP 3: $\boldsymbol{\alpha_i \sim N(0, 0.02^2), i = 1, \ldots, 200; \alpha_i = 0, i = 201, \ldots, 2000}$

In this DGP, 200 of the 2000 funds have zero alphas and the remaining funds have alphas generated randomly from $N(0, 0.02^2)$ *with half being set to be positive and half being set to be negative. This DGP lies somewhere between DGP 1 and DGP 2. Given 90% of all funds either outperform or underperform their benchmark, we should expect that realized alphas are above their bootstrapped counterparts with likelihoods close to one at top percentiles and below their bootstrapped values with likelihoods close to zero at bottom percentiles.*

Panels A and B of Table 6 show the results based on alpha estimates and their t-statistics under two error distributions $N(0, 0.08^2)$ and $N(0,1)$, respectively. In our simulation, fund returns are generated from the market model (1), in which, 200 funds have zero alphas and the remaining funds have alphas drawn randomly from $N(0, 0.02^2)$ with half being set to be positive and half being set to be negative. We simulate market returns from $N(0.08, 0.15^2)$ and βs from $U[0.5, 2]$. The errors in Panels A and B are generated randomly from $N(0, 0.08^2)$ and $N(0, 0.64^2)$, respectively.

In Panel A, the approach works very well. At both top and bottom percentiles, there are significant differences between realized alphas and their bootstrapped values. Likelihoods are uniformly zero at percentiles below 10th and 100% at percentiles above 80th.

However, as we increase the standard deviation of errors to 0.64, the approach starts losing its power. As shown in Panel B of Table 6, at top percentiles, the realized and

bootstrapped alphas are virtually indistinguishable. At bottom percentiles, the likelihoods are above 10%, indicating no significant difference between realized and bootstrapped alphas. The reason for this is the same as for DGP 2.

In untabulated results which are available from the authors upon request, we have found qualitatively similar inferences drawn from results based on t-statistics, qualitatively similar results for uniform, t and NIG errors.

In untabulated results which are available from the authors upon request, the variance decompositions and Kolmogorov-Smirnov tests show strong evidence for a difference between the distributions of realized alphas and bootstrapped alphas for DGPs 2 and 3.

In summary, given reasonably large errors, the bootstrap evaluation approach works well when a significant fraction of funds is assumed to either outperform or underperform the market benchmark. The likelihood-based tests used by Fama and French (2010) among others, will lead to correct inferences on fund skills, based on either alpha estimates or their t-statistics. However, as the standard deviation of errors in the market model increases to 0.64, Fama and French (2010)'s interpretation of likelihoods can result in problematic inferences on fund skills.

5. Results for the Four-Factor Model from Carhart (1997)

In this section, we use the four-factor model (2) to generate fund returns and examine how the bootstrap evaluation approach performs when the four-factor model is used as the performance benchmark.

We obtain monthly data for r_{mt}, SMB_t, HML_t,and MOM_t from French's online data library. The data cover the period from January 1984 to September 2006, the same as in Fama and French (2010). Details on simulation settings are described in Sections 3.1 and 3.2.

We report the results for DGP 1 only since the bootstrap evaluation approach typically fails in DGP 1 as shown in Section 4. Results based on t-statistics are reported whenever necessary.

5.1 Percentiles and Likelihoods Based on the Four-Factor Model

Table 7 displays realized and bootstrapped alphas and likelihoods at the selected percentiles for DGP 1 under four different distributions of errors in the four-factor model: uniform, normal, t and NIG in Panels A, B, C and D, respectively. The results are based on independent sampling of fund residuals.

We simulate βs from the uniform distribution on $[0.5, 2]$ and sample errors from one of the four distributions: uniform, normal, t and NIG distributions. Parameters of the four distributions are set so that errors have zero mean and annual standard deviation of 0.08. Specifically, the uniform distribution lies on the interval $[-0.04,0.04]$, the normal distribution has zero mean and annual standard deviation of 0.08, the t distribution has a degree of freedom of 3, and the NIG distribution has tail heaviness λ=10, location μ=0, asymmetry parameter θ=0 and scale parameter δ=0.0053.

In Panel A, we find that the approach works very well. At both tails, the realized and bootstrapped alphas are quite close to each other. From the first to fifth percentiles, bootstrapped values are below their realized counterparts in a minimum of 45.84% to a maximum of 86.58% of the 5000 bootstraps, indicating no significant difference between them. A similar conclusion can be drawn from the top percentiles.

However, in Panel B with normal errors, the approach loses its power. At the 97th percentile, the bootstrapped value is 3.33, smaller than the realized value 3.57 in 99.08% of the 5000 bootstraps, which, according to Fama and French (2010), should be accepted as evidence of superior skills. In addition, likelihoods are above 70% at most of the selected

percentiles and above 90% at the 98th and 99th percentiles. It seems that realized alphas at these percentiles tend to be upward biased.

The approach is also invalid in Panel C, in which errors are sampled from the *t* distribution. At the 95th percentile, the realized alpha is 3.03, above its bootstrapped value 2.84. The likelihood 98.46% indicates superior skills among top funds.

In Panel D, errors are sampled from the NIG distribution. We find that likelihoods are below 50% at most percentiles; however, there is no strong evidence of skills at both tails. Therefore, the approach gives correct inferences on skills.

In summary, similar to what we find from the market model, the bootstrap evaluation approach can produce incorrect inferences on fund skills when all funds are endowed with zero alphas. Therefore, in this case, likelihoods cannot be relied on to make statistical inferences on fund skills.

5.2 Results from Variance Decompositions and Kolmogorov-Smirnov tests

Figure 2 displays PDFs and CDFs of realized and bootstrapped alphas under four different error distributions. We see that under each of the four error distributions, both PDF and CDF (dashed line) for realized alphas almost coincide with those of bootstrapped alphas (solid line), which indicates no skills.

Table 8 reports the kernel density estimates of mean and variance of alpha estimates and their t-statistics and the results for Kolmogorov-Smirnov tests. The variance ratio is computed as the ratio of kernel estimates of variances of bootstrapped and realized alphas. The Kolmogorov-Smirnov test is used to test for difference between realized and bootstrapped alphas.

In Panel A, we see that regardless of the error distribution, the variance ratio is very close to one. Therefore, the variation of realized alphas can be largely attributed to random

errors. The variance ratios of the t-statistics are also close to one as shown in Panel B. In addition, the Kolmogorov-Smirnov tests in both Panels A and B demonstrate that we cannot reject the null of no difference between realized and bootstrapped alphas.

In contrast to percentiles and likelihoods, the variance decomposition and the Kolmogorov-Smirnov test can help us to make correct inferences on fund skills when all funds have zero alphas.

6. Consequences of Violations of Regression Assumptions

The bootstrap evaluation approach is partially motivated by the fact that the classical OLS regression assumptions are often violated in earlier fund evaluation studies. In this section, we investigate the possible consequences brought by the violations of classical OLS regression assumptions: including serial correlation, GARCH effect and contemporaneous correlation among errors; multicollinearity in factor returns; and omitted factor.

6.1 Serial Correlation among Errors

We assume that errors in (1) follow an AR(1) process

$$\varepsilon_{it} = \phi\varepsilon_{it-1} + u_{it}, i = 1, 2, \dots, 2000 \quad (14)$$

where $u_{it} \sim i.i.d\ N(0, \sigma_u^2)$.

The mean, variance and the kth order correlation of the AR(1) process are $E(\varepsilon_{it}) = 0$, $\sigma_\varepsilon^2 = \sigma_u^2/(1-\phi^2)$, and $\rho(k) = \phi^k$, respectively. In our simulation, we set ϕ=0.10 and σ_u=0.0796 so that the annualized standard deviation σ_ε of errors is 0.08. Other parameters are specified the same as in (i) and (ii) in Section 3.2.

Panels A, B and C of Table 9 report the results for DGP 1 based on alpha estimates, standard t-statistics, and Newey-West t-statistics (Newey and West (1987)), respectively. [12]

We find that the approach does not work. In Panel A, we observe strong evidence of superior skills at top percentiles and evidence of inferior skills at bottom percentiles. At top five percentiles, 100% of the 5000 bootstrapped values are below their realized counterparts, indicating of superior skills. In contrast, at bottom five percentiles, all the likelihoods are zero, indicating inferior skills. The same conclusion can be drawn from results based on t-statistics in Panel B. In Panel C, the realized t-statistics are shrunk compared to those in Panel B. However, the Newey-West t-statistics remain revealing significant superior skills among top-performing funds and inferior skills among bottom-performing funds even though likelihoods are different from 100% at top and zero at bottom.

From regression analysis, we know that serial correlations among errors affect the efficiency of OLS estimators, which implies the increase in the standard deviation of realized alphas. Compared to realized alphas in Panel B of Table 3, we do find that the dispersion of realized alphas has increased. Furthermore, for a positive serial correlation as in our case, the standard errors of coefficient estimates will be underestimated. As a result, their t-statistics will be inflated. We find that the dispersion in the standard t-statistics in Panel B of Table 9 increases relative to that of the t-statistics in Panel B of Table 4. While Newey-West t-statistics make some improvements by shrinking the standard t-statistics, it is not enough to

[12] To compute Newey-West serial-correlation-robust standard errors, we set the lag length to be five in accordance with the current practice. In the current practice, the commonly used lag length is the smallest integer greater than or equal to $T^{1/4}$, see Greene (2011). The sample length T is 273 in our simulation.

produce correct inferences on skills.[13] In addition, bootstrapped alphas and their t-statistics are not affected by serial correlations among errors.

Interestingly, in untabulated results which are available from the authors upon request, we find that the bootstrap approach works well for DGPs 2 and 3.

6.2 GARCH Effects among Errors

Empirical studies show that financial returns exhibit volatility clustering. To examine its impact on the performance of the bootstrap evaluation approach, we use the following GARCH (1, 1) process to model errors in the market model (1)

$$\varepsilon_{it} = \sigma_{it} u_{it},\ i = 1, 2, \ldots, 2000 \tag{15}$$

$$\sigma_{it}^2 = \omega + \theta \varepsilon_{it-1}^2 + \gamma \sigma_{it-1}^2 \tag{16}$$

where $u_{it} \sim i.i.d\ N(0,1)$, $\omega > 0, \gamma > 0, \theta > 0$, and $\theta + \gamma < 1$.

Under the assumption of covariance stationarity, the unconditional variance of errors is $\sigma_\varepsilon^2 = \omega/(1-\theta-\gamma)$. In our simulation, we set ω= 0.000027, θ=0.05 and γ=0.90 so that the annualized standard deviation σ_ε is equal to 0.08. Other parameters are specified as in (i) and (ii) in Section 3.2.

Panels A, B, and C of Table 10 report the results based on t-statistics for DGPs 1, 2, and 3, respectively.[14] In Panel A, at top percentiles, the realized alphas are below their

[13] From results unreported, we find other things being equal, as serial correlations rise from zero, the inflation in t-statistics of realized alphas will exacerbate. On the other hand, a negative correlation leads to the shrinkage of realized alphas and their t-statistics. As the correlation decreases from zero, the shrinkage will aggravate. In both cases, the cross-section of bootstrapped alphas stays approximately the same as in the absence of serial correlations. For negative serial correlations, the pattern of likelihoods is also reversed: virtually zero at upper percentiles while very close to 100% at lower percentiles.

bootstrapped counterparts with likelihoods below 5%, which is evidence of inferior skills among top-performing funds and inconsistent with our assumption of no skills. However, this approach works well in DGPs 2 and 3 as shown in Panels B and C.[15]

6.3 Contemporaneous Correlations among Errors

To examine the effect of contemporaneous correlations among errors on the bootstrap evaluation approach, we simulate errors in the market model (1) from a multivariate normal distribution with zero means, annual standard deviations 0.08 and an identical and constant correlation across all funds 0.1. Other parameters are specified the same as in (i) and (ii) in Section 3.2.

Panels A, B, and C of Table 11 report results based on t-statistics for DGPs 1, 2 and 3 under two different bootstrap schemes: an independent sampling of fund residuals (*Bstrap1*) and joint sampling of fund residuals (*Bstrap2*).

We find that while the joint sampling scheme leads to the shrinkage of the distribution of bootstrapped alphas and thus an improvement in some situations, it can also produce incorrect inferences on skills. First, comparing *Bstrap1* and *Bstrap2* in Panels A, B, and C, we find the joint sampling scheme shrinks bootstrapped alphas in all these three DGPs. Second, likelihoods in Column 3 of Panel A indicate that there is inferior skills among bottom funds. In contrast, likelihoods based on a joint sampling of fund residuals in Column 5 show that the approach works very well. In Panel B, the approach also works well

[14] Results based on alpha estimates and Newey-West t-statistics provide no additional insights and thus are not reported here.

[15] From results unreported, we find that similar to what we observe in Section 4.2, as the ratio of the standard deviations of errors and actual alphas increases, it will be more and more difficult to differentiate the t-statistics of realized alphas and bootstrapped alphas.

regardless of the bootstrap scheme. However, in Panel C, both schemes fail. At all the selected percentiles, the bootstrapped values based on independent sampling (*Bstrap1*) are above their realized counterparts. Furthermore, likelihoods in Column 3 of Panel C are zero at all the selected percentiles, indicating of no superior skills among top-performing funds, in contradiction to our assumption. The bootstrapped values based on joint sampling (*Bstrap2*) are also uniformly above their realized counterparts; however, likelihoods in Column 5 of Panel C indicate of neither superior nor inferior skills.

Even though our assumption of a constant and identical correlation of 0.1 between all funds is far away from reality, it does highlight how sensitive the bootstrap evaluation approach is to contemporaneous correlations among errors. Moreover, even with the knowledge of contemporaneous correlations among errors, the use of joint sampling of fund residuals is not foolproof as shown in DGP 3.

6.4 Multicollinearity in Factor Returns

To examine the impact of multicollinearity on the performance of the bootstrap evaluation approach, we simulate factor returns in the four-factor model (2) from a multivariate normal distribution with mean vector μ=[0.8, 0.8, 0.8, 0.8], standard deviations σ=[0.15, 0.17, 0.19, 0.21] and correlation coefficient ρ=0.5 among all the factors. Other parameters are simulated in the same way as in (i) and (iii) in Section 3.2 except that for each fund, we have to generate four factor loadings from $U[0.5,\ 2]$.

Table 12 reports percentiles and likelihoods for DGPs 1, 2 and 3 based on t-statistics of alpha estimates.

We find that the approach works in all these three DGPs although the t-statistics at virtually all percentiles in DGPs 2 and 3 are slightly shrunk toward zero compared to those in Panel A of Table 5 and those in Panel A of Table 6, respectively. We know that in regression

analysis, the standard errors of coefficient estimates tend to be large in the presence of multicollinearity in explanatory variables. Consequently, the t-statistics tend to be smaller, which explains the shrinkage in the t-statistics of realized alphas in Panels B and C.

6.5 Omitted Factor

In this subsection, we examine how the omitted factor from the benchmark model will affect the performance of the bootstrap evaluation approach. We use Ferson and Schadt's (1996) conditional beta model to generate fund returns, but the market model (1) as the benchmark fund performance evaluation model.

Ferson and Schadt's (1996) Conditional Beta Model

In market model (1), fund betas are measured as averages over the evaluation period without regard to the state of financial markets. Now we extend the market model (1) to the conditional beta model developed by Ferson and Schadt (1996)

$$r_{i,t+1} = \alpha_i + \beta_{1i} r_{m,t+1} + \beta' [r_{m,t+1} \otimes Z_t] + \varepsilon_{i,t} \tag{17}$$

where $r_{i,t+1}$ is the excess return of fund i, Z_t is the vector of lagged conditioning variables in demeaned form. The symbol $\otimes$ denotes the Kronecker product. This model amounts to a multifactor model by treating $r_{m,t+1} \otimes Z_t$ as a vector of additional factors, see Jagannathan and Wang (1996). It can also be written as a market model with a time-varying beta $\beta(Z_t)$

$$r_{i,t+1} = \alpha_i + \beta(Z_t)\, r_{m,t+1} + \varepsilon_{i,t} \tag{18}$$

where $\beta(Z_t) = \beta_{1i} + \beta' Z_t$.

In our simulation, we assume that there is only one state variable Z_t, representing some predictor of r_m, and simulate Z_t from $N(0, 0.15^2)$. We generate $r_{m,t+1}$ from

$$r_{m,t+1} = \gamma Z_t + u_{t+1} \tag{19}$$

where $u_{t+1} \sim N(0, 0.08^2)$ and γ=0.5. We simulate funds' exposures to the factor $r_{m,t+1} \otimes Z_t$ from $U[0.5, 2]$. Other parameters are specified the same as described in Sections 3.1 and 3.2.

As before, the market model (1) is used for performance evaluation. The resulting problem is equivalent to omitting the factor $r_{m,t+1} \otimes Z_t$. In regression analysis, if some explanatory variable is omitted, the coefficient estimates will be biased upward or downward, depending on the correlation between the omitted factor and other explanatory variables. Moreover, the standard error of coefficient estimates will be biased positively. The biases on the coefficient estimate can either cancel or reinforce the bias in the standard error. Therefore, the net effect on t-statistics is not clear.

Simulation Results

Figure 3 plots the kernel density estimates of probability density functions (PDFs) of realized (dashed line) and bootstrapped (solid line) alphas for DGP 1: all funds are endowed with zero alphas.

We see that the PDF of realized alphas lies to the right of that for bootstrapped alphas, implying that at all the percentiles, realized alphas should be larger than those bootstrapped. This is at odds with the assumption that all funds have zero alphas.

Panels A, B and C of Table 13 show the percentiles and likelihoods (%<*Rea*) for realized (*Rea*) and bootstrapped (*Bstrap*) alphas at the selected percentiles (*Pct*) for DGPs 1, 2 and 3, respectively. We also report the percentiles for realized alphas (*Rea_No*) when the factor $r_{m,t+1} \otimes Z_t$ in (17) is taken into account for estimation.

In Panel A, all funds are endowed with zero alphas. We find that the approach is invalid. First, realized alphas (*Rea*) are overestimated relative to their counterparts (*Rea_No*) when $r_{m,t+1} \otimes Z_t$ is taken into account. Consequently, at all the selected percentiles, the realized values are larger than their bootstrapped counterparts with likelihoods virtually equal

to 100%. The reason is, given the positive value of γ, the omission of $r_{m,t+1} \otimes Z_t$ in the benchmark model leads to a positive bias in alpha estimates. As a result, the entire distribution of realized alphas shifts to the right as illustrated in Figure 3.[16]

In Panel B, all funds have alphas randomly generated from $N(0, 0.02^2)$. We find that realized alphas are larger than those when $r_{m,t+1} \otimes Z_t$ is taken into account. However, the approach works since the difference between realized alphas and their bootstrapped values are sufficiently large.[17] In Panel C, the approach remains valid.

Results based on t-statistics are similar and thus are not reported here.

To sum up, the bootstrap evaluation approach can lose its power in the presence of the following issues: serial correlation, GARCH effect and contemporaneous correlation among errors; multicollinearity in factor returns; and omitted factor especially when all funds are endowed with zero alphas. For example, when errors in the market model follow an AR(1) process and all funds are endowed with zero alphas, likelihoods indicate strong evidence of superior skills among top-performing funds and of inferior skills among bottom-ranking funds. This is true regardless of using alpha estimates or their t-statistics. When errors follow GARCH (1,1) and all funds are endowed with zero alphas, likelihoods can result in evidence of inferior skills among top-performing funds. When all funds are endowed with non-zero alphas, likelihoods can usually lead to correct inferences on skills at extreme tails.

[16] From unreported results, we find that when γ is negative, realized alphas will be underestimated relative to their counterparts when $r_{m,t+1} \otimes Z_t$ is taken into account. The PDF of realized alphas will lie to the left of the PDF of bootstrapped alphas. Likelihoods in Case 1 will be virtually zero at the selected percentiles.

[17] However, from unreported results, we find that as γ in (19) rises, the distribution of realized alphas when $r_{m,t+1} \otimes Z_t$ is omitted will get closer and closer to that of bootstrapped alphas. Nevertheless, other things being equal, at extreme tails, the method remains valid even if γ reaches the level 0.9.

7. Performance of Alternative Bootstrap Schemes

The independent sampling of fund residuals disregards the correlation and dependence among errors and factor returns. In this section, we consider three alternative bootstrap schemes: (1) joint sampling of fund residuals (2) joint sampling of fund residuals and factor returns and (3) joint sampling of fund residuals and factor returns using block bootstrap. It is worth noting that different bootstrap schemes influence statistical inferences on skills by affecting bootstrapped alphas but not realized alphas. Likelihoods will change in response to the change in bootstrapped alphas.

For comparison, we use the four-factor model (2) as the benchmark and use the same factor loadings and errors generated in Table 7, Section 5.

Panels A, B and C of Table 14 report realized and bootstrapped alphas at the selected percentiles for DGP 1 under the alternative bootstrap schemes (1), (2) and (3), respectively.

In Panel A, we see that the percentiles of bootstrapped alphas are slightly shrunk toward zero compared to those obtained from independent sampling of fund residuals in Panel A of Table 7. This is true for all the four error distributions. For example, in the case of uniform errors, at the first, second and third percentiles, the likelihoods are respectively 43.80%, 49.88% and 71.30%, down from their respective values 45.84% 55.34%, and 86.58% in Table 7; at the 97th, 98th and 99th percentiles, likelihoods are respectively 64.10%, 77.38% and 79.42%, down from 68.28%, 86.34% and 85.12%. Therefore, compared to the independent sampling of fund residuals, the joint sampling of fund residuals tends to shrink the distribution of bootstrapped alphas and thus likelihoods at tails.

Results in Panel B are based on jointly sampling fund residuals and factor returns. We find that likelihoods at lower percentiles slightly increase while those at upper percentiles slightly decrease. More importantly, these likelihoods lead to correct inferences on fund

skills. For example, under uniform errors, at the first, second and third percentiles, likelihoods increase to 48.80%, 55.26% and 71.80% from 43.80%, 49.88% and 71.30% in Panel A. At the 97th, 98th and 99th percentiles, likelihoods are respectively 55.12%, 66.56% and 68.46%, further down from 64.10%, 77.38% and 79.42 %, their values in Panel A. For normal errors, likelihoods at all the percentiles are below 90% with the maximum 87.44% occurring at the fifth percentile. We also see reduced dispersions across likelihoods under t and NIG errors. Under t errors, the likelihood ranges from 13.20% to 81.36% in Panel B while it ranges from 7.30% to 91.06% in Panel A. Under NIG errors, the minimum likelihood increase to 8.82% from 4.62% in Panel A while the maximum is roughly the same as in Panel A.

All the prior bootstrap schemes ignore the serial dependence in fund residuals and factor returns. The block bootstrap considers this issue. In our simulation, there are virtually zero serial correlations among errors because we generate them from $i.i.d.$ distributions. The major concern is due to the market data on factor returns. While the block length can be arbitrarily chosen, we consider the block lengths of {3, 4, 6, 10, 12} months, at which lags, the returns on at least one of the factors have autocorrelations above 10%.

Table 15 displays the autocorrelations of factor returns and the results of Ljung-Box Q-tests up to 15 lags. At some lags, the autocorrelations are not negligible: -11% and 11% at lags 4 and 11 for the *Market Factor*, -13% at lag 3 for *Size(SMB)*, 10% for *Value (HML)* at lag 1, and 13% at lag 6 for *Momentum*. At lag 12, the autocorrelation for *Momentum* is 19% and the p-value for the Q-test indicates that we can reject the null of no serial correlation at the significance level of 0.01. Given these numbers, it is interesting to know whether the block bootstrap can improve the likelihood-based statistical inferences on fund skills.

Panel C of Table 14 shows the results for a block length of 10.[18] We find that depending on the error distribution, likelihoods at tails can be either increase or decrease relative to those in Panel B. As a result, likelihoods lead to correct inferences on fund skills: all funds have zero alphas. Under the uniform errors, at the third, fourth and fifth percentiles, likelihoods are reduced to 54.82%, 54.78% and 54.20% in Panel C from 71.80%, 71.50% and 70.72% in Panel B. Under the normal errors, at the third, fourth and fifth percentiles, likelihoods decrease to 47.04%, 61.72% and 65.16% in Panel C from 62.34%, 83.06% and 87.44% in Panel B. At top percentiles, there is not much change in both cases. In contrast, for t and NIG errors, significant changes occur at top percentiles. For example, under NIG errors, at the 95th to 99th percentiles, likelihoods are respectively 53.24%, 52.16%, 54.86%, 52.86% and 42.78%, up from 42.96%, 39.84%, 45.46%, 41.36% and 26.28% in Panel B.

In summary, we find that while the independent sampling of fund residuals is not reliable, both the joint sampling of fund residuals and factor returns and block bootstrap are valid when alpha estimates are used. The block bootstrap scheme seems to be more reliable given the dependence structure in the market data on factor returns.[19]

We also conduct similar analyses for these three alternative schemes using t-statistics. From results unreported here, we find that the block bootstrap remains valid and reliable when t-statistics are used; however, the joint sampling of fund residuals and factor returns can be misleading. In this sense, t-statistics have no advantage over alpha estimates in inferences on fund skills. Our findings also support Fama and French (2010)'s claim that the

[18] For other block lengths, similar results are found and thus are not reported here.

[19] We also check the variance decomposition and Kolmogrov-Smirnov tests for these three alternative bootstrap schemes and find that they allow us to make correct inferences on skill when all funds are endowed with zero alphas.

bootstrap scheme used by Kosowski et al. (2006) may well be one of the reasons behind their strong results.

8. Concluding Remarks

We provide a Monte Carlo simulation analysis of the validity and reliability of a bootstrap evaluation approach, which has recently been applied to identifying skills in a large strand literature but with striking different results. In general, we find that the bootstrap approach can lead to problematic inferences about fund skills and thus cast doubts on its recent applications. On the one hand, our analyses show that when all funds are endowed with zero alphas, the approach can lead to the problematic inference that there exists superior skills among top funds. On the other hand, if the funds are endowed with non-zero alphas, it can effectively establish the existence of skills, superior or bad. However, if errors in the market model exhibit extremely high dispersion relative to fund alphas, the approach can also fail to detect the existence of skills.

The major problem with this approach lies in the inappropriate use of what Fama and French (2010) call "likelihoods" in testing for difference between realized and bootstrapped alphas at the selected percentiles. Fama and French (2010), among others, treat likelihoods as if they were the p-values of statistical tests. However, this is not correct even though it does not always lead to problematic inferences. Fama and French (2010)'s likelihood at the selected percentile is obtained based on historical returns of mutual funds, which is just one realization from the underlying distribution of fund returns.

The use of t-statistics rather than alpha estimates in the bootstrap evaluation does not guarantee any advantage in terms of validity and reliability of the bootstrap evaluation approach. On the contrary, because of the strong robustness property of the t-statistics, a

focus on t-statistics alone may hide the potential issues that can be detected from examining alpha estimates sometimes.

Alternative bootstrap schemes that correct for correlations and preserve data dependence structure tend to shrink the cross-sectional distribution of bootstrapped alphas and can result in improvement in statistical inferences on fund skills. This is in support of Fama and French (2010)'s claim that different bootstrap schemes may explain the difference between their findings on US mutual fund performance and those by Kosowski et al. (2006). However, these alternative bootstrap schemes are not foolproof either.

The bootstrap evaluation approach is partially motivated by the fact that the classical OLS regression assumptions are often violated in earlier fund evaluation studies. Consequently, we investigate the possible consequences brought by the violations of classical OLS regression assumptions: including serial correlation, GARCH effect and contemporaneous correlation among errors; multicollinearity in factor returns; and omitted factor. We find that, in the presence of most problems (i.e., serial correlation, GARCH effect and contemporaneous correlation among errors; and multicollinearity in factor returns), the bootstrap evaluation approach may falsely identify fund skills when all funds are endowed with zero alphas. The consequence is particularly serious in the presence of some omitted factor, as the bootstrap evaluation approach may mistake superior (inferior) skills for inferior (superior) due to underestimating (overestimating) realized alphas via OLS.

Overall, we find that the bootstrap evaluation approach hinges critically on the appropriateness of using likelihoods as the p-values for statistical inferences on fund skills. Theoretically, we could obtain the empirical distribution of likelihoods at each percentile by generating a large number of paths of fund returns, and compare the likelihoods obtained from historical returns against its empirical distribution to test for the difference between

realized and bootstrapped alphas. However, this would be computationally intensive and unnecessary. Once the distribution of the realized alphas is obtained, a simple variance decompositions and Kolmogorov-Smirnov tests can come to rescue when the bootstrap evaluation approach fails. When all funds are endowed with zero alphas, even if the bootstrap evaluation approach tends to predict skills, the variance ratio will be very close to one and the Kolmogorov-Smirnov test will not reject the null of no difference between the distributions of bootstrapped and realized alphas. For researchers and practitioners who are interested in fund performance evaluation, this paper provides guidance and intuition regarding the possible deficiency of using the bootstrap approach in various ways, and insights for improving the existing bootstrap schemes as well as alternative non-bootstrap performance evaluation approaches (e.g., Chen et al., 2017; Ferson and Chen 2017; Harvery and Liu, 2017).

Table 1. Sensitivity Analysis of Realized and Bootstrapped Alphas (DGP 1)

The table reports annualized realized (Rea) and bootstrapped (Bstrp) alphas in percentages and likelihoods (%<Rea) at selected percentiles (Pct) for DGP 1, in which, all funds are endowed with zero alphas. Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) refers to the proportion of bootstrapped alphas out of 5000 bootstraps that fall below their realized counterparts at the selected percentiles. In our simulation, fund returns are generated from the market model (1), in which market returns and errors are drawn randomly from $N(0.08, 0.15^2)$ and $N(0, 0.08^2)$, respectively. Otherwise identical, Panels A, B and C have βs simulated randomly from $U[0.5, 2]$, $U[0.5, 3.5]$ and $N(1.25, 0.43^2)$, respectively. Otherwise identical, Panels A and D are based on two different realizations of errors from $N(0, 0.08^2)$.

	Panel A: Baseline			**Panel B: β_i ~ U[0.5, 3.5]**			**Panel C: β_i ~ N[1.25, 0.43^2]**			**Panel D: $\varepsilon_{i,t}$ ~ N[0, 0.08^2]**		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-3.75	-3.90	86.80	-3.75	-3.90	84.70	-3.75	-3.89	85.40	-3.95	-3.90	37.50
2	-3.42	-3.44	58.90	-3.42	-3.44	56.50	-3.42	-3.43	54.70	-3.48	-3.44	35.80
3	-2.97	-3.15	97.00	-2.97	-3.15	97.70	-2.97	-3.14	98.10	-3.19	-3.15	33.40
4	-2.78	-2.93	95.80	-2.78	-2.93	95.90	-2.78	-2.93	96.00	-2.88	-2.93	72.10
5	-2.63	-2.75	95.10	-2.63	-2.75	94.00	-2.63	-2.75	94.70	-2.73	-2.75	62.80
10	-2.07	-2.14	87.90	-2.07	-2.14	87.10	-2.07	-2.14	89.30	-2.18	-2.14	25.20
20	-1.28	-1.41	98.90	-1.28	-1.40	98.70	-1.28	-1.40	98.70	-1.43	-1.40	34.40
30	-0.81	-0.87	89.00	-0.81	-0.87	88.80	-0.81	-0.88	89.30	-0.89	-0.88	39.50
40	-0.40	-0.42	62.80	-0.40	-0.42	63.30	-0.40	-0.42	65.10	-0.46	-0.42	19.20
50	-0.02	0.00	31.50	-0.02	0.00	34.20	-0.02	0.00	36.70	0.02	0.00	67.90
60	0.35	0.43	6.90	0.35	0.42	5.80	0.35	0.42	6.30	0.46	0.42	78.40
70	0.89	0.88	61.00	0.89	0.88	60.40	0.89	0.88	62.00	0.88	0.87	53.80
80	1.42	1.40	61.30	1.42	1.41	56.00	1.42	1.41	59.80	1.38	1.40	30.40
90	2.15	2.14	55.90	2.15	2.14	54.30	2.15	2.14	53.40	2.05	2.14	8.30
95	2.76	2.75	55.00	2.76	2.75	54.60	2.76	2.75	54.30	2.65	2.75	10.30
96	2.94	2.93	57.80	2.94	2.93	55.80	2.94	2.93	56.90	2.85	2.93	19.30
97	3.15	3.14	50.50	3.15	3.15	49.60	3.15	3.15	51.00	3.09	3.15	25.70
98	3.57	3.43	88.90	3.57	3.44	86.70	3.57	3.44	88.70	3.35	3.44	21.80
99	4.04	3.89	84.80	4.04	3.89	84.80	4.04	3.89	85.20	3.74	3.90	12.00

Table 2. Variance Ratios and Kolmogorov-Smirnov Tests (DGP 1)

This table presents the kernel density estimates of means and variances of realized (Rea) and bootstrapped (Bstrap) alphas and the results of Kolmogorov-Smirnov tests for difference between realized and bootstrapped alphas. The variance ratio is the ratio of the variance of bootstrapped alphas to that of realized alphas. In our simulation, fund returns are generated from the market model (1), in which all funds are endowed with zero alphas, market returns are randomly drawn from $N(0.08, 0.15^2)$, βs are simulated from $U[0.5, 2]$ and errors are randomly sampled from one of the four distributions: uniform, normal, t and NIG. Parameters of the four distributions are set so that errors have zero mean and annual standard deviation 0.08. Specifically, the uniform distribution lies on $[-0.04,0.04]$, the t distribution has a degree of freedom 3 and the NIG distribution has tail heaviness λ=10, location μ=0, asymmetry parameter θ=0 and scale parameter δ=0.0053. For Kolmogorov-Smirnov tests, "0" indicates of no rejection of the null of no difference in the two distributions.

	Uniform		**Normal**		**Student-t**		**NIG**	
	Rea	Bstrap	Rea	Bstrap	Rea	Bstrap	Rea	Bstrap
Mean	0.10	0.00	0.04	0.00	0.00	0.00	-0.07	0.00
Variance	2.90	2.95	2.88	2.94	2.97	2.94	2.97	2.93
Variance ratio	0.98		0.98		1.01		1.02	
Kolmogorov-Smirnov	0		0		0		0	

Table 3. Percentiles of Alpha Estimates (DGP 1)

The table shows annualized realized (Rea) and bootstrapped (Bstrap) alphas in percentages and likelihoods (%<Rea) at the selected percentiles (Pct) for DGP 1 under four different error distributions: uniform, normal, t and NIG in Panels A, B, C and D, respectively. Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of bootstrapped alphas out of 5000 bootstraps that fall below their realized counterparts at the selected percentiles. In our simulation, fund returns are generated from the market model (1), in which all funds are endowed with zero alphas, market returns are randomly generated from $N(0.08, 0.15^2)$, βs are simulated from $U[0.5, 2]$ and errors in Panels A, B, C and D are sampled from uniform, normal, t and NIG distributions, respectively. Parameters of the four distributions are set so that errors have zero mean and annual standard deviation 0.08. Specifically, the uniform distribution lies on $[-0.04, 0.04]$, the t distribution has a degree of freedom 3 and the NIG distribution has tail heaviness λ=10, location μ=0, asymmetry parameter θ=0 and scale parameter δ=0.0053.

	Panel A: Uniform			**Panel B: Normal**			**Panel C: Student-t**			**Panel D: NIG**		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-3.99	-3.89	22.20	-3.95	-3.90	37.50	-3.89	-3.98	69.60	-4.23	-4.08	20.20
2	-3.55	-3.44	15.30	-3.48	-3.44	35.80	-3.48	-3.42	30.80	-3.63	-3.50	14.40
3	-3.20	-3.14	26.00	-3.19	-3.15	33.40	-3.30	-3.10	2.10	-3.16	-3.15	45.60
4	-3.00	-2.93	18.80	-2.88	-2.93	72.10	-3.07	-2.87	1.50	-2.92	-2.90	42.30
5	-2.86	-2.75	8.00	-2.73	-2.75	62.80	-2.87	-2.68	1.60	-2.73	-2.71	38.50
10	-2.19	-2.14	23.60	-2.18	-2.14	25.20	-2.22	-2.06	0.60	-2.11	-2.06	26.80
20	-1.41	-1.41	45.70	-1.43	-1.40	34.40	-1.41	-1.34	9.80	-1.32	-1.33	57.40
30	-0.87	-0.88	58.40	-0.89	-0.88	39.50	-0.89	-0.83	12.50	-0.82	-0.82	54.70
40	-0.39	-0.42	78.20	-0.46	-0.42	19.20	-0.46	-0.40	7.40	-0.34	-0.40	90.30
50	0.03	0.00	74.10	0.02	0.00	67.90	-0.05	0.00	14.30	0.05	0.00	87.80
60	0.45	0.42	69.50	0.46	0.42	78.40	0.35	0.40	14.50	0.44	0.40	86.10
70	0.86	0.87	37.00	0.88	0.87	53.80	0.77	0.83	11.90	0.88	0.82	90.10
80	1.41	1.41	49.00	1.38	1.40	30.40	1.35	1.34	62.20	1.38	1.33	85.00
90	2.13	2.14	45.60	2.05	2.14	8.30	2.01	2.06	20.90	2.20	2.07	98.50
95	2.87	2.75	92.50	2.65	2.75	10.30	2.62	2.68	25.50	2.88	2.71	97.10
96	3.07	2.92	96.30	2.85	2.93	19.30	2.79	2.86	20.00	3.05	2.91	92.70
97	3.31	3.15	95.80	3.09	3.15	25.70	3.03	3.10	25.40	3.31	3.16	91.10
98	3.63	3.43	96.70	3.35	3.44	21.80	3.35	3.42	29.50	3.65	3.51	87.00
99	4.15	3.89	96.00	3.74	3.90	12.00	3.84	3.96	25.70	4.25	4.09	84.40

Table 4. Percentiles of t-Statistics of Alpha Estimates (DGP 1)

The table shows the t-statistics of realized alphas (Rea) and bootstrapped alphas (Bstrap) and likelihoods (%<Rea) at selected percentiles (Pct) for DGP 1 under four different distributions for errors: uniform, normal, t and NIG in Panels A, B, C and D, respectively. Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of t-statistics of bootstrapped alphas out of 5000 bootstraps that fall below those of realized alphas at the selected percentiles. In our simulation, fund returns are generated from the market model (1), in which all funds are endowed with zero alphas, market returns are simulated from $N(0.08, 0.15^2)$, βs are generated randomly from $U[0.5, 2]$ and errors in Panels A, B, C and D are sampled from the four distributions: uniform, normal, t and NIG, respectively. Parameters of the four distributions are set so that the errors have zero mean and annual standard deviation 0.08. Specifically, the uniform distribution lies on $[-0.04, 0.04]$, the t distribution has a degree of freedom 3 and the NIG distribution has tail heaviness λ=10, location μ=0, asymmetry parameter θ=0 and scale parameter δ=0.0053.

	Panel A: Uniform			**Panel B: Normal**			**Panel C: Student-t**			**Panel D: NIG**		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-2.37	-2.34	35.70	-2.31	-2.34	64.90	-2.48	-2.40	18.00	-2.22	-2.44	99.20
2	-2.13	-2.06	14.30	-2.09	-2.06	35.70	-2.19	-2.10	9.60	-2.03	-2.13	92.10
3	-1.95	-1.89	12.70	-1.92	-1.89	29.60	-2.03	-1.92	2.80	-1.87	-1.94	88.90
4	-1.79	-1.76	25.40	-1.74	-1.76	66.70	-1.88	-1.78	3.40	-1.74	-1.80	88.00
5	-1.70	-1.65	12.00	-1.62	-1.65	72.40	-1.76	-1.67	3.90	-1.65	-1.69	81.90
10	-1.31	-1.28	24.60	-1.30	-1.28	30.70	-1.35	-1.30	11.30	-1.27	-1.31	87.50
20	-0.85	-0.84	40.60	-0.86	-0.84	30.70	-0.89	-0.85	14.00	-0.81	-0.86	93.50
30	-0.51	-0.53	68.50	-0.53	-0.52	38.90	-0.56	-0.53	19.30	-0.52	-0.54	71.00
40	-0.23	-0.25	79.70	-0.28	-0.25	19.50	-0.28	-0.26	17.80	-0.21	-0.26	94.10
50	0.02	0.00	74.30	0.01	0.00	67.10	-0.03	0.00	14.50	0.03	0.00	87.80
60	0.26	0.25	66.00	0.28	0.25	79.00	0.22	0.26	10.30	0.30	0.26	89.90
70	0.51	0.52	28.20	0.52	0.52	43.70	0.49	0.53	8.60	0.57	0.54	85.70
80	0.84	0.84	41.70	0.82	0.84	25.00	0.85	0.85	48.10	0.87	0.86	60.80
90	1.28	1.28	47.30	1.22	1.28	5.90	1.24	1.30	8.40	1.35	1.31	83.90
95	1.70	1.65	85.70	1.59	1.65	11.20	1.61	1.67	11.30	1.77	1.68	96.20
96	1.83	1.76	92.50	1.68	1.76	5.70	1.69	1.78	6.10	1.86	1.79	90.90
97	1.98	1.89	94.10	1.85	1.89	25.40	1.79	1.91	1.90	2.00	1.93	89.20
98	2.16	2.06	93.40	1.98	2.06	10.20	1.92	2.09	0.70	2.13	2.12	61.70
99	2.48	2.34	93.70	2.24	2.34	13.60	2.26	2.39	7.60	2.48	2.42	74.40

Table 5. Percentiles of Alpha Estimates and Their t-Statistics (DGP 2)

The table reports annualized realized (Rea) and bootstrapped (Bstrap) alphas (in percentages), their t-statistics and likelihoods (%<Rea) at selected percentiles (Pct) for DGP 2 under two different error distributions $N(0, 0.08^2)$ and $N(0,1)$ in Panels A and B, respectively. Details on how to obtain percentiles for alphas and their t-statistics are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of alpha estimates or their t-statistics out of the 5000 bootstraps that fall below their realized counterparts at the selected percentiles. In our simulation, fund returns are generated from the market model (1), in which, fund alphas are simulated from $N(0, 0.02^2)$, market returns are generated randomly from $N(0.08, 0.15^2)$ and βs are drawn from $U[0.5, 2]$. Errors in Panels A and B are generated randomly from $N(0, 0.08^2)$ and $N(0,1)$, respectively, which is the only difference between them.

	Panel A: Normal (0, 0.08)						**Panel B: Normal (0, 1)**					
	Alpha (%)			t-statistics			Alpha (%)			t-statistics		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-17.01	-3.90	0	-9.94	-2.34	0	-51.27	-48.63	7.10	-2.45	-2.34	8.70
2	-14.95	-3.44	0	-9.06	-2.06	0	-44.38	-42.89	13.20	-2.11	-2.06	27.90
3	-13.41	-3.15	0	-8.24	-1.89	0	-41.40	-39.28	4.20	-1.98	-1.89	4.70
4	-12.79	-2.93	0	-7.63	-1.75	0	-38.27	-36.56	5.60	-1.85	-1.76	3.70
5	-11.86	-2.75	0	-7.19	-1.65	0	-36.60	-34.36	1.40	-1.77	-1.65	0.60
10	-9.07	-2.14	0	-5.38	-1.28	0	-28.58	-26.71	1.00	-1.37	-1.28	1.30
20	-5.67	-1.40	0	-3.41	-0.84	0	-18.89	-17.52	1.80	-0.90	-0.84	3.40
30	-3.54	-0.87	0	-2.11	-0.52	0	-11.00	-10.93	47.00	-0.53	-0.53	47.50
40	-1.50	-0.42	0	-0.90	-0.25	0	-4.93	-5.29	74.20	-0.23	-0.25	78.00
50	0.37	0.00	100	0.21	0.00	100	0.04	-0.02	53.90	0.00	0.00	54.00
60	2.05	0.42	100	1.26	0.25	100	5.30	5.25	54.00	0.26	0.25	60.40
70	4.01	0.87	100	2.42	0.53	100	11.88	10.90	95.10	0.56	0.52	88.90
80	6.14	1.40	100	3.67	0.84	100	18.99	17.52	98.80	0.90	0.84	96.50
90	9.43	2.14	100	5.56	1.28	100	28.07	26.71	94.60	1.33	1.28	88.20
95	11.71	2.75	100	7.09	1.65	100	35.46	34.33	86.50	1.70	1.65	84.80
96	12.51	2.92	100	7.49	1.75	100	37.09	36.54	69.30	1.77	1.76	58.40
97	13.47	3.14	100	8.07	1.89	100	39.34	39.29	51.00	1.91	1.89	63.50
98	14.71	3.44	100	9.02	2.06	100	43.36	42.94	63.40	2.14	2.06	89.70
99	16.99	3.89	100	10.47	2.34	100	51.18	48.67	92.20	2.38	2.34	68.50

Table 6. Percentiles of Alpha Estimates and Their t-Statistics (DGP 3)

The table reports annualized realized (Rea) and bootstrapped (Bstrap) alphas in percentages, their t-statistics and likelihoods (%<Rea) at selected percentiles (Pct) for DGP 3 under two different error distributions $N(0, 0.08^2)$ and $N(0, 0.64^2)$ in Panels A and B, respectively. Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of alpha estimates or their t-statistics out of 5000 bootstraps that fall below their realized counterparts at the selected percentiles. In our simulation, fund returns are generated from the market model (1). 90% of funds have zero alphas and the remaining 10% have alphas generated randomly from $N(0, 0.02^2)$ with half being positive and half being negative. We simulate market returns from $N(0.08, 0.15^2)$ and βs from $U[0.5, 2]$. The only difference between Panels A and B lies in the errors. They are generated randomly from $N(0, 0.08^2)$ in Panel A and from $N(0, 0.64^2)$ in Panel B.

	Panel A: Normal (0, 0.08)						**Panel B: Normal (0, 0.64)**					
	Alpha (%)			t-statistics			Alpha (%)			t-statistics		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-9.92	-3.89	0.00	-5.89	-2.34	0.00	-31.38	-31.18	42.30	-2.41	-2.34	21.60
2	-6.49	-3.43	0.00	-3.84	-2.06	0.00	-28.19	-27.44	19.80	-2.11	-2.06	25.00
3	-4.83	-3.15	0.00	-2.85	-1.89	0.00	-26.12	-25.13	10.90	-1.93	-1.89	22.70
4	-3.84	-2.93	0.00	-2.35	-1.76	0.00	-24.08	-23.39	15.60	-1.81	-1.76	14.80
5	-3.56	-2.75	0.00	-2.08	-1.65	0.00	-22.28	-21.99	31.60	-1.64	-1.65	59.60
10	-2.38	-2.14	0.00	-1.43	-1.28	0.00	-17.23	-17.12	41.60	-1.28	-1.29	52.50
20	-1.46	-1.40	15.40	-0.87	-0.84	19.10	-11.58	-11.23	21.50	-0.86	-0.84	31.70
30	-0.90	-0.87	27.20	-0.54	-0.52	28.20	-7.21	-7.00	29.10	-0.53	-0.53	36.70
40	-0.44	-0.42	34.70	-0.26	-0.25	39.70	-3.45	-3.38	41.40	-0.26	-0.25	39.20
50	0.00	0.00	51.80	0.00	0.00	51.80	-0.03	0.00	45.80	0.00	0.00	45.90
60	0.38	0.42	16.20	0.22	0.25	14.80	3.32	3.38	43.90	0.24	0.25	36.60
70	0.98	0.87	98.50	0.59	0.52	98.60	6.92	7.00	40.50	0.52	0.53	37.70
80	1.60	1.40	100.00	0.96	0.84	100.00	11.25	11.22	51.20	0.84	0.84	42.60
90	2.42	2.14	100.00	1.43	1.28	100.00	18.03	17.11	96.10	1.33	1.29	88.80
95	3.54	2.75	100.00	2.10	1.65	100.00	23.02	21.97	94.20	1.73	1.65	95.50
96	3.93	2.93	100.00	2.37	1.76	100.00	24.53	23.40	94.60	1.81	1.76	84.10
97	4.98	3.14	100.00	2.89	1.89	100.00	25.88	25.14	83.80	1.96	1.89	90.40
98	6.67	3.43	100.00	3.92	2.06	100.00	28.80	27.47	92.60	2.16	2.06	91.60
99	10.16	3.89	100.00	6.18	2.34	100.00	32.53	31.11	89.20	2.50	2.34	97.10

Table 7. Percentiles of Alpha Estimates (the Four-Factor Model, DGP 1)

This table shows annualized realized (Rea) and bootstrapped (Bstrap) alphas in percentages and likelihoods (%<Rea) at selected percentiles (Pct) for DGP 1: all funds are endowed with zero alphas under four different error distributions: uniform, normal, t and NIG. Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of bootstrapped alphas out of the 5000 bootstraps that fall below their realized counterparts at the selected percentiles. Factor returns over the period January 1984 to September 2006 are obtained from French's online data library. We simulate βs from the uniform distribution on $[0.5, 2]$ and sample errors from one of the four different distributions: uniform, normal, t and NIG distributions. Parameters of the four distributions are set so that errors have zero mean and annual standard deviation of 0.08. Specifically, the uniform distribution lies on $[-0.04, 0.04]$, the normal distribution has zero mean and annual standard deviation 0.08, the t distribution has a degree of freedom 3 and the NIG distribution has tail heaviness λ=10, location μ=0, asymmetry parameter θ=0 and scale parameter δ=0.0053.

	Panel A: Uniform [-0.04, 0.04]			**Panel B: N(0, 0.08^2)**			**Panel C: Student-t**			**Panel D: NIG**		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-4.12	-4.11	45.84	-4.35	-4.13	7.34	-4.26	-4.20	34.64	-4.24	-4.43	81.32
2	-3.61	-3.63	55.34	-3.59	-3.64	65.46	-3.71	-3.63	26.90	-3.83	-3.75	30.20
3	-3.21	-3.32	86.58	-3.27	-3.33	70.56	-3.30	-3.29	44.60	-3.39	-3.36	39.86
4	-2.99	-3.09	86.40	-2.92	-3.10	97.84	-3.11	-3.04	22.74	-3.18	-3.08	17.26
5	-2.82	-2.91	85.54	-2.72	-2.91	99.00	-2.96	-2.84	9.76	-3.02	-2.86	5.16
10	-2.22	-2.27	74.50	-2.16	-2.26	94.22	-2.34	-2.18	1.08	-2.35	-2.15	0.34
20	-1.45	-1.49	75.56	-1.41	-1.49	91.48	-1.51	-1.42	4.06	-1.49	-1.37	2.08
30	-0.91	-0.93	63.02	-0.86	-0.93	90.46	-0.96	-0.88	5.48	-0.94	-0.84	2.78
40	-0.48	-0.45	28.36	-0.40	-0.45	80.80	-0.49	-0.42	10.56	-0.44	-0.40	17.22
50	0.00	0.00	48.86	0.02	0.00	67.20	0.01	0.00	55.04	0.00	0.00	48.44
60	0.41	0.45	24.06	0.52	0.45	93.48	0.46	0.42	78.96	0.37	0.41	21.00
70	0.90	0.93	29.90	0.96	0.92	77.24	0.91	0.88	74.74	0.83	0.85	35.06
80	1.49	1.49	49.68	1.51	1.49	67.30	1.50	1.42	93.72	1.40	1.38	67.58
90	2.33	2.26	84.02	2.27	2.26	52.68	2.32	2.18	97.92	2.20	2.15	77.14
95	2.98	2.91	79.26	2.96	2.91	73.46	3.03	2.84	98.46	2.83	2.85	40.36
96	3.09	3.09	48.60	3.20	3.10	88.66	3.12	3.03	81.66	3.03	3.07	36.60
97	3.37	3.32	68.28	3.57	3.33	99.08	3.30	3.28	58.38	3.34	3.35	47.18
98	3.76	3.63	86.34	3.81	3.64	93.88	3.55	3.62	30.30	3.71	3.74	42.78
99	4.27	4.11	85.12	4.34	4.12	93.04	4.01	4.19	15.60	4.24	4.41	20.94

Table 8. Variance Ratios and Kolmogorov-Smirnov Tests (Four-Factor, DGP 1)

This table presents the kernel density estimates of mean and variance of the bootstrapped and realized alphas and their t-statistics, variance ratios and the results for Kolmogorov-Smirnov tests for DGP 1. Details on how to obtain realized alphas, bootstrapped alphas and their t-statistics are given in Sections 2.2 and 2.3. The variance ratio is the ratio of the variance of bootstrapped alphas (or their t-statistics) to that of realized alphas (or their t-statistics). Errors are sampled from one of the four different distributions: uniform, normal, t and NIG distributions. Parameters of the four distributions are set so that the errors have a zero mean and an annual standard deviation of 0.08. Specifically, the uniform distribution lies on $[-0.04,0.04]$, the t distribution has a degree of freedom 3 and the NIG distribution has tail heaviness λ=10, location μ=0, asymmetry parameter θ=0 and scale parameter δ=0.0053. For Kolmogorov-Smirnov tests, "0" indicates of no rejection of the null of no difference in the two distributions.

Model 1:	**Uniform**		**Normal**		**Student-t**		**NIG**	
Independent	Rea	Bstrap	Rea	Bstrap	Rea	Bstrap	Rea	Bstrap
Panel A: alpha estimates								
Mean	0.01	0.00	0.05	0.00	-0.01	0.00	-0.05	0.00
Variance	3.31	3.29	3.26	3.29	3.43	3.22	3.31	3.28
Variance ratio	1.01		0.99		1.06		1.01	
Kolmogorov-Smirnov	0		0		0		0	
Panel B: t-statistics of alpha								
Mean	0.01	0.00	0.03	0.00	-0.01	0.00	-0.03	0.00
Variance	1.05	1.06	1.03	1.06	1.10	1.09	1.05	1.14
Variance ratio	0.99		0.97		1.02		0.92	
Kolmogorov-Smirnov	0		0		0		0	

Table 9. Percentiles of Alpha Estimates and t-Statistics (AR(1) Errors, DGP 1)

The table reports annualized realized (Rea) and bootstrapped (Bstrap) alphas in percentages, their standard and Newey-West t-statistics and likelihoods (%<Rea) at selected percentiles (Pct) in Panels A, B and C, respectively. Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood is the proportion of alpha estimates or their t-statistics out of the 5000 bootstraps that fall below their realized counterparts at the selected percentiles. In our simulation, fund returns are generated from the market model (1) with errors simulated from the AR(1) model (14). We set fund alphas to be zero, simulate market returns from $N(0.08, 0.15^2)$ and sample βs from $U[0.5, 2]$. We let ϕ=0.10 and σ_u=0.0796 in (14) so that the annualized standard deviation of errors σ_ε is 0.08. We use the lag length of 5 to compute Newey-West serial-correlation-robust standard errors.

	Panel A: $\hat{\alpha}$			**Panel B: $t_{\hat{\alpha}}$**			**Panel C: $t_{\hat{\alpha}}$ (Newey-West)**		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-4.40	-3.91	0.00	-2.70	-2.34	0.00	-2.48	-2.40	18.00
2	-3.90	-3.44	0.00	-2.36	-2.06	0.00	-2.25	-2.11	1.70
3	-3.56	-3.15	0.00	-2.15	-1.89	0.00	-2.03	-1.93	5.70
4	-3.31	-2.93	0.00	-2.00	-1.76	0.00	-1.90	-1.79	2.60
5	-3.06	-2.75	0.00	-1.81	-1.65	0.00	-1.78	-1.68	2.10
10	-2.41	-2.14	0.00	-1.44	-1.28	0.00	-1.34	-1.30	15.20
20	-1.52	-1.41	1.20	-0.91	-0.84	1.70	-0.84	-0.85	68.50
30	-0.96	-0.88	3.70	-0.57	-0.53	6.70	-0.53	-0.53	50.80
40	-0.48	-0.42	10.70	-0.29	-0.25	11.30	-0.27	-0.26	36.60
50	-0.02	0.00	30.90	-0.01	0.00	30.20	-0.01	0.00	30.60
60	0.49	0.43	90.70	0.28	0.26	82.30	0.26	0.26	57.90
70	0.99	0.88	98.90	0.59	0.53	98.60	0.55	0.53	75.60
80	1.63	1.41	100.00	0.96	0.84	100.00	0.88	0.85	76.80
90	2.43	2.14	100.00	1.44	1.28	100.00	1.33	1.30	77.10
95	3.18	2.75	100.00	1.91	1.65	100.00	1.78	1.68	97.50
96	3.40	2.93	100.00	2.02	1.75	100.00	1.88	1.79	95.30
97	3.63	3.15	100.00	2.16	1.89	100.00	2.06	1.93	98.60
98	4.01	3.44	100.00	2.37	2.06	100.00	2.23	2.11	95.90
99	4.41	3.90	100.00	2.57	2.33	99.30	2.44	2.40	69.20

Table 10. Percentiles of t-Statistics (GARCH(1, 1) Errors)

The table shows the t-statistics of realized (Rea) and bootstrapped (Bstrap) alphas and likelihoods (%<Rea) at selected percentiles (Pct) for DGPs 1, 2, and 3 in Panels A, B and C, respectively. Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of the t-statistics out of the 5000 bootstraps that fall below those of realized alphas at the selected percentiles. In Panel A, all funds are endowed with zero alphas. In Panel B, all the funds have alphas generated randomly from $N(0, 0.02^2)$. In Panel C, 10% of the funds have alphas generated randomly from $N(0, 0.02^2)$ while the remaining funds have zero alphas. In our simulation, fund returns are generated from the market model (1) with market returns are simulated from $N(0.08, 0.15^2)$, βs are sampled from $U[0.5, 2]$ and errors are generated from GARCH(1, 1) described in (15) and (16). We set ω= 0.000027, θ=0.05 and γ=0.90 so that the annualized standard deviation σ_ε is equal to 0.08.

	Panel A: Case 1			**Panel B: Case 2**			**Panel C: Case 3**		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-2.31	-2.33	58.30	-11.36	-2.34	0.00	-5.69	-2.34	0.00
2	-2.09	-2.06	27.20	-10.07	-2.06	0.00	-4.32	-2.06	0.00
3	-1.89	-1.88	47.40	-9.19	-1.89	0.00	-2.86	-1.89	0.00
4	-1.77	-1.75	35.10	-8.76	-1.76	0.00	-2.41	-1.76	0.00
5	-1.68	-1.65	23.90	-8.18	-1.65	0.00	-2.17	-1.65	0.00
10	-1.29	-1.28	37.10	-5.88	-1.29	0.00	-1.52	-1.29	0.00
20	-0.86	-0.84	32.50	-3.77	-0.84	0.00	-0.93	-0.84	0.50
30	-0.54	-0.52	34.20	-2.19	-0.53	0.00	-0.58	-0.53	3.00
40	-0.26	-0.25	37.40	-1.03	-0.25	0.00	-0.30	-0.25	6.10
50	-0.03	0.00	14.00	0.06	0.00	98.60	-0.04	0.00	9.50
60	0.23	0.25	20.30	1.30	0.25	100.00	0.25	0.25	45.10
70	0.49	0.53	14.40	2.43	0.52	100.00	0.53	0.53	60.90
80	0.80	0.84	7.60	3.75	0.84	100.00	0.87	0.84	77.30
90	1.21	1.28	1.30	5.83	1.29	100.00	1.42	1.28	100.00
95	1.53	1.65	0.50	7.57	1.65	100.00	1.97	1.65	100.00
96	1.61	1.76	0.00	8.11	1.76	100.00	2.18	1.76	100.00
97	1.73	1.89	0.10	8.77	1.89	100.00	2.74	1.89	100.00
98	1.92	2.06	1.00	9.42	2.06	100.00	3.90	2.06	100.00
99	2.17	2.34	1.90	10.40	2.34	100.00	5.74	2.34	100.00

Table 11. Percentiles of t-Statistics (Contemporaneously Correlated Errors)

The table shows the t-statistics of realized (Rea) and bootstrapped (Bstrap) alphas and likelihoods (%<Rea) at selected percentiles (Pct) for DGPs 1, 2, and 3 in Panels A, B and C, respectively. We use two bootstrap schemes to obtain bootstrapped t-statistics: independent sampling of fund residuals (Bstrap1) and joint sampling of fund residuals (Bstrap2). Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of the t-statistics out of the 5000 bootstraps that fall below those of realized alphas at the selected percentiles. In Panel A, all funds are endowed with zero alphas. In Panel B, all the funds have alphas generated randomly from $N(0, 0.02^2)$. In Panel C, 10% of the funds have alphas generated randomly from $N(0, 0.02^2)$ while the remaining funds have zero alphas. In our simulation, fund returns are generated from the market model (1), in which, market returns are simulated from $N(0.08, 0.15^2)$, βs are sampled from $U[0.5, 2]$ and errors are drawn from the multivariate normal distribution with zero means, annual standard deviations 0.08 and a constant correlation across all the funds 0.1.

	Panel A: Case 1					**Panel B: Case 2**					**Panel C: Case 3**				
Pct	Rea	Bstrap1	%<Rea	Bstrap2	%<Rea	Rea	Bstrap1	%<Rea	Bstrap2	%<Rea	Rea	Bstrap1	%<Rea	Bstrap2	%<Rea
1	-2.43	-2.34	12.60	-2.21	24.90	-3.70	-2.34	0	-2.20	0	-2.67	-2.34	0	-2.22	8.50
2	-2.18	-2.06	3.00	-1.95	23.80	-3.35	-2.06	0	-1.94	0	-2.34	-2.06	0	-1.96	12.20
3	-2.03	-1.89	0.70	-1.78	22.90	-3.13	-1.89	0	-1.78	0	-2.11	-1.89	0	-1.79	16.20
4	-1.91	-1.76	0.10	-1.66	22.30	-2.94	-1.76	0	-1.66	0	-2.01	-1.76	0	-1.67	14.60
5	-1.85	-1.65	0	-1.56	19.20	-2.73	-1.65	0	-1.56	0	-1.89	-1.65	0	-1.57	15.70
10	-1.48	-1.28	0	-1.21	21.40	-2.24	-1.28	0	-1.21	0.10	-1.53	-1.28	0	-1.22	18.00
20	-1.10	-0.84	0	-0.79	17.70	-1.61	-0.84	0	-0.79	0.40	-1.12	-0.84	0	-0.81	16.80
30	-0.79	-0.52	0	-0.49	18.10	-1.08	-0.53	0	-0.49	2.30	-0.82	-0.52	0	-0.50	16.70
40	-0.51	-0.25	0	-0.23	18.90	-0.72	-0.25	0	-0.23	5.70	-0.53	-0.25	0	-0.25	20.20
50	-0.28	0.00	0	0.01	18.40	-0.32	0.00	0	0.01	15.70	-0.27	0.00	0	-0.01	21.00
60	-0.04	0.25	0	0.26	18.10	0.09	0.25	0	0.25	29.20	-0.02	0.25	0	0.24	21.90
70	0.20	0.53	0	0.51	17.20	0.54	0.53	69.20	0.51	53.70	0.23	0.53	0	0.49	21.70
80	0.47	0.84	0	0.82	14.60	1.00	0.84	100.00	0.81	72.10	0.53	0.84	0	0.79	21.60
90	0.94	1.29	0	1.24	18.00	1.65	1.29	100.00	1.23	91.00	0.99	1.28	0	1.21	25.60
95	1.29	1.65	0	1.59	18.70	2.08	1.65	100.00	1.57	94.00	1.36	1.65	0	1.56	28.80
96	1.39	1.76	0	1.69	19.10	2.31	1.76	100.00	1.67	97.50	1.47	1.76	0	1.66	29.80
97	1.50	1.89	0	1.81	18.40	2.50	1.89	100.00	1.80	98.50	1.57	1.89	0	1.78	26.70
98	1.66	2.07	0	1.97	17.30	2.67	2.06	100.00	1.96	98.80	1.76	2.06	0	1.94	30.50
99	1.92	2.35	0	2.23	18.20	3.04	2.34	100.00	2.22	99.60	2.07	2.34	0	2.20	36.30

Table 12. Percentiles of t-Statistics When Factor Returns are Correlated

The table reports t-statistics of realized (Rea) and bootstrapped (Bstrap) alphas and likelihoods (%<Rea) at the selected percentiles (Pct) for DGPs 1, 2, and 3 in Panels A, B and C, respectively. Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of the t-statistics out of the 5000 bootstraps that fall below those of realized alphas at the selected percentiles. In Panel A all funds are endowed with zero alphas. In Panel B, the funds have alphas generated randomly from $N(0, 0.02^2)$. In Panel C, 10% of the funds have alphas generated randomly from $N(0, 0.02^2)$ while the remaining funds have zero alphas. In our simulation, fund returns are generated from the four-factor model (2), in which, factor loadings βs are drawn randomly from $U[0.5, 2]$, errors are simulated from $N(0, 0.08^2)$ and factor returns are drawn randomly from a multivariate normal distribution with mean vector μ=[0.8, 0.8, 0.8, 0.8], standard deviations σ=[0.15, 0.17, 0.19, 0.21] and correlation coefficient ρ=0.5 among all the factors.

	Panel A: Case 1			**Panel B: Case 2**			**Panel C: Case 3**		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
1	-2.33	-2.34	53.90	-9.61	-2.34	0.00	-5.71	-2.34	0.00
2	-2.04	-2.06	62.80	-8.48	-2.06	0.00	-4.06	-2.06	0.00
3	-1.93	-1.89	24.60	-7.79	-1.89	0.00	-2.78	-1.89	0.00
4	-1.76	-1.76	50.80	-7.39	-1.76	0.00	-2.28	-1.76	0.00
5	-1.64	-1.65	59.00	-6.98	-1.65	0.00	-2.06	-1.65	0.00
10	-1.28	-1.29	55.20	-5.56	-1.28	0.00	-1.49	-1.28	0.00
20	-0.84	-0.84	47.40	-3.55	-0.84	0.00	-0.93	-0.84	0.50
30	-0.54	-0.53	27.90	-2.32	-0.53	0.00	-0.61	-0.53	0.40
40	-0.31	-0.25	2.20	-1.03	-0.25	0.00	-0.33	-0.25	0.20
50	-0.04	0.00	6.20	0.05	0.00	97.40	-0.06	0.00	1.90
60	0.21	0.25	7.40	1.11	0.25	100.00	0.22	0.25	15.50
70	0.46	0.53	1.60	2.34	0.53	100.00	0.51	0.52	28.20
80	0.80	0.84	6.30	3.68	0.84	100.00	0.87	0.84	73.50
90	1.23	1.28	6.40	5.26	1.29	100.00	1.42	1.28	100.00
95	1.63	1.65	32.80	6.95	1.65	100.00	2.00	1.65	100.00
96	1.75	1.76	43.50	7.37	1.76	100.00	2.19	1.76	100.00
97	1.87	1.89	38.90	8.01	1.89	100.00	2.49	1.89	100.00
98	2.02	2.07	25.10	8.48	2.07	100.00	3.13	2.06	100.00
99	2.22	2.34	6.70	9.29	2.34	100.00	5.28	2.34	100.00

Table 13. Percentiles of Alpha Estimates in the Presence of Omitted Factor

The table reports annualized realized (Rea) and bootstrapped (Bstrap) alphas (in percentages) and likelihoods (%<Rea) at selected percentiles (Pct). Whereas Rea_No is based on realized alphas when $r_{m,t+1} \otimes Z_t$ in (17) is taken into account, Rea is based on realized alphas when $r_{m,t+1} \otimes Z_t$ is omitted from the benchmark model. Btrap are the average values of alpha estimates at the selected percentiles from 5000 bootstraps based on independent sampling of fund residuals. The likelihood (%<Rea) is the proportion of the t-statistics out of 5000 bootstraps that fall below those of realized alphas at the selected percentiles. In Panel A, all funds are endowed with zero alphas. In Panel B, the funds have alphas randomly generated from $N(0, 0.02^2)$. In Panel C, 10% of the funds have alphas randomly drawn from $N(0, 0.02^2)$ while the remaining funds have zero alphas. In our simulation, excess returns on the market factor $r_{m,t+1}$ are generated from (19), in which, γ is set to be 0.5, u_{t+1} and Z_t are simulated from $N(0, 0.08^2)$ and $N(0, 0.15^2)$, respectively. Fund returns are generated from the conditional beta model (17), in which, factor loadings βs and errors $\varepsilon_{i,t}$ are randomly drawn from $U[0.5, 2]$ and $N(0, 0.08^2)$, respectively.

	Panel A: Case 1				**Panel B: Case 2**				**Panel C: Case 3**			
Pct	Rea_No	Rea	Bstrap	%<Rea	Rea_No	Rea	Bstrap	%<Rea	Rea_No	Rea	Bstrap	%<Rea
1	-3.96	-3.70	-3.89	91.50	-16.50	-16.43	-3.90	0	-9.96	-9.72	-3.91	0
2	-3.46	-3.18	-3.44	99.40	-14.42	-14.28	-3.44	0	-6.32	-6.10	-3.44	0
3	-3.16	-2.96	-3.15	98.50	-13.49	-13.22	-3.15	0	-4.50	-4.20	-3.15	0
4	-3.01	-2.77	-2.93	96.30	-12.59	-12.41	-2.93	0	-3.91	-3.64	-2.93	0
5	-2.83	-2.53	-2.75	99.50	-11.84	-11.57	-2.75	0	-3.47	-3.23	-2.75	0
10	-2.16	-1.92	-2.14	100	-9.49	-9.18	-2.14	0	-2.51	-2.24	-2.14	6.10
20	-1.35	-1.16	-1.41	100	-6.28	-5.97	-1.40	0	-1.55	-1.33	-1.41	93.30
30	-0.86	-0.61	-0.88	100	-3.97	-3.68	-0.88	0	-0.95	-0.70	-0.88	100
40	-0.39	-0.15	-0.42	100	-1.96	-1.69	-0.42	0	-0.44	-0.22	-0.42	100
50	0.03	0.27	0.00	100	0.04	0.17	0.00	100	0.03	0.27	0.00	100
60	0.48	0.68	0.42	100	1.80	2.13	0.42	100	0.53	0.73	0.42	100
70	0.92	1.15	0.88	100	3.84	4.09	0.88	100	1.02	1.27	0.88	100
80	1.51	1.69	1.41	100	6.27	6.42	1.41	100	1.65	1.88	1.41	100
90	2.20	2.37	2.14	100	9.57	9.85	2.14	100	2.45	2.66	2.14	100
95	2.74	2.95	2.75	99.40	12.01	12.06	2.75	100	3.55	3.76	2.75	100
96	2.98	3.18	2.93	99.90	12.67	12.79	2.93	100	3.92	4.17	2.93	100
97	3.21	3.42	3.15	99.70	14.09	14.21	3.15	100	4.62	4.74	3.15	100
98	3.45	3.70	3.44	99.20	14.98	15.27	3.44	100	6.65	6.54	3.44	100
99	3.91	4.11	3.91	92.40	16.95	17.16	3.90	100	10.25	10.37	3.90	100

Table 14. Percentiles of Alpha Estimates (Other Bootstrap Schemes, DGP 1)

The table reports annualized realized (Rea) and bootstrapped (Bstrap) alphas (in percentages) and likelihoods (%<Rea) at the selected percentiles (Pct) under four error distributions: uniform, normal, t and NIG. We consider three alternative bootstrap schemes: joint sampling of fund residuals (Panel A), joint sampling of fund residuals and factor returns (Panel B) and joint sampling of fund residuals and factor returns using block bootstrap (Panel C). Details on how to obtain percentiles and likelihoods are given in Sections 2.2 and 2.3. The likelihood (%<Rea) is the proportion of bootstrapped alphas out of 5000 bootstraps that fall below their realized values at the selected percentiles. All funds are endowed with zero alphas and the four-factor model is used as the benchmark. Factor returns from January 1984 to September 2006 come from French's online data library. βs and errors are the same as those in Table 7.

	Uniform [-0.04, 0.04]			$N(0, 0.08^2)$			Student-t			NIG		
Pct	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea	Rea	Bstrap	%<Rea
Panel A: Joint sampling of fund residuals only												
1	-4.12	-4.09	43.80	-4.35	-4.11	15.30	-4.26	-4.17	36.20	-4.24	-4.42	72.72
2	-3.61	-3.61	49.88	-3.59	-3.63	56.42	-3.71	-3.61	31.22	-3.83	-3.75	35.18
3	-3.21	-3.31	71.30	-3.27	-3.32	58.94	-3.30	-3.28	43.78	-3.39	-3.36	43.24
4	-2.99	-3.08	71.30	-2.92	-3.09	86.58	-3.11	-3.03	28.94	-3.18	-3.08	28.02
5	-2.82	-2.90	69.68	-2.72	-2.91	90.50	-2.96	-2.83	19.74	-3.02	-2.86	15.02
10	-2.22	-2.26	62.56	-2.16	-2.26	81.28	-2.34	-2.18	7.30	-2.35	-2.16	4.62
20	-1.45	-1.49	66.40	-1.41	-1.49	82.28	-1.51	-1.42	11.22	-1.49	-1.38	7.08
30	-0.91	-0.93	58.46	-0.86	-0.93	85.56	-0.96	-0.88	8.32	-0.94	-0.84	5.90
40	-0.48	-0.45	29.24	-0.40	-0.45	79.80	-0.49	-0.42	10.90	-0.44	-0.40	20.08
50	0.00	0.00	48.32	0.02	0.00	66.92	0.01	0.00	55.44	0.00	0.00	46.96
60	0.41	0.45	25.44	0.52	0.45	90.34	0.46	0.42	77.98	0.37	0.41	22.06
70	0.90	0.93	35.36	0.96	0.93	71.38	0.91	0.88	71.54	0.83	0.85	37.94
80	1.49	1.49	51.74	1.51	1.49	60.32	1.50	1.42	86.68	1.40	1.38	61.92
90	2.33	2.26	73.90	2.27	2.27	51.20	2.32	2.18	89.58	2.20	2.15	68.26
95	2.98	2.90	69.92	2.96	2.91	63.66	3.03	2.83	91.06	2.83	2.85	43.68
96	3.09	3.08	52.28	3.20	3.10	75.54	3.12	3.03	73.80	3.03	3.07	41.58
97	3.37	3.31	64.10	3.57	3.33	91.44	3.30	3.27	57.54	3.34	3.35	48.04
98	3.76	3.62	77.38	3.81	3.63	82.72	3.55	3.61	40.50	3.71	3.74	45.94
99	4.27	4.09	79.42	4.34	4.11	84.20	4.01	4.17	26.92	4.24	4.41	29.92
Panle B: Joint sampling of fund residuals and factor returns												
1	-4.12	-4.14	48.80	-4.35	-4.15	24.34	-4.26	-4.24	44.66	-4.24	-4.48	73.76
2	-3.61	-3.66	55.26	-3.59	-3.67	60.12	-3.71	-3.66	39.64	-3.83	-3.79	40.98
3	-3.21	-3.35	71.80	-3.27	-3.36	62.34	-3.30	-3.31	48.62	-3.39	-3.38	46.82
4	-2.99	-3.12	71.50	-2.92	-3.13	83.06	-3.11	-3.06	37.80	-3.18	-3.10	33.30
5	-2.82	-2.93	70.72	-2.72	-2.94	87.44	-2.96	-2.86	29.58	-3.02	-2.87	21.88
10	-2.22	-2.29	64.68	-2.16	-2.29	80.70	-2.34	-2.20	15.76	-2.35	-2.16	9.92
20	-1.45	-1.50	68.18	-1.41	-1.50	82.06	-1.51	-1.43	18.46	-1.49	-1.37	11.30
30	-0.91	-0.93	60.32	-0.86	-0.94	84.70	-0.96	-0.89	13.72	-0.94	-0.84	8.82
40	-0.48	-0.45	30.66	-0.40	-0.45	79.52	-0.49	-0.43	13.20	-0.44	-0.40	21.56
50	0.00	0.00	46.40	0.02	0.00	64.04	0.01	0.00	55.62	0.00	0.00	48.60
60	0.41	0.46	22.34	0.52	0.46	85.46	0.46	0.43	72.52	0.37	0.41	24.18
70	0.90	0.94	31.66	0.96	0.94	63.56	0.91	0.89	64.42	0.83	0.85	41.14
80	1.49	1.51	44.26	1.51	1.51	53.84	1.50	1.43	76.78	1.40	1.38	61.94
90	2.33	2.29	62.70	2.27	2.29	46.00	2.32	2.20	79.10	2.20	2.16	64.24
95	2.98	2.94	60.00	2.96	2.94	55.16	3.03	2.86	81.36	2.83	2.87	42.96
96	3.09	3.13	45.60	3.20	3.13	64.78	3.12	3.06	64.04	3.03	3.09	39.84
97	3.37	3.36	55.12	3.57	3.36	81.42	3.30	3.31	51.60	3.34	3.38	45.46
98	3.76	3.66	66.56	3.81	3.67	72.20	3.55	3.65	37.70	3.71	3.78	41.36
99	4.27	4.14	68.46	4.34	4.16	73.48	4.01	4.23	26.48	4.24	4.48	26.28
Panle C: Block sampling of fund residuals and factor returns (Block length=10)												
1	-4.12	-4.05	41.42	-4.35	-4.06	26.44	-4.26	-4.15	39.06	-4.24	-4.40	57.38
2	-3.61	-3.59	44.46	-3.59	-3.58	46.34	-3.71	-3.58	36.80	-3.83	-3.72	37.44
3	-3.21	-3.29	54.82	-3.27	-3.28	47.04	-3.30	-3.25	42.22	-3.39	-3.32	41.20
4	-2.99	-3.06	54.78	-2.92	-3.06	61.72	-3.11	-3.00	35.28	-3.18	-3.04	33.06
5	-2.82	-2.88	54.20	-2.72	-2.87	65.16	-2.96	-2.80	30.62	-3.02	-2.83	26.76
10	-2.22	-2.24	49.80	-2.16	-2.24	58.64	-2.34	-2.16	21.44	-2.35	-2.12	17.88
20	-1.45	-1.47	53.00	-1.41	-1.47	60.52	-1.51	-1.40	23.70	-1.49	-1.35	18.44
30	-0.91	-0.92	49.78	-0.86	-0.91	66.44	-0.96	-0.87	18.90	-0.94	-0.83	13.90
40	-0.48	-0.44	28.80	-0.40	-0.44	69.26	-0.49	-0.42	15.14	-0.44	-0.40	21.42
50	0.00	0.00	47.34	0.02	0.00	66.80	0.01	0.00	54.68	0.00	0.00	47.00
60	0.41	0.45	33.22	0.52	0.44	85.02	0.46	0.42	74.72	0.37	0.40	35.66
70	0.90	0.92	46.12	0.96	0.92	66.92	0.91	0.87	66.26	0.83	0.83	50.32
80	1.49	1.48	55.62	1.51	1.47	61.46	1.50	1.40	74.16	1.40	1.35	63.30
90	2.33	2.25	64.90	2.27	2.24	56.96	2.32	2.16	75.72	2.20	2.12	65.16
95	2.98	2.88	63.68	2.96	2.88	62.22	3.03	2.80	76.96	2.83	2.82	53.24
96	3.09	3.07	55.34	3.20	3.06	67.62	3.12	3.00	66.24	3.03	3.04	52.16
97	3.37	3.29	60.94	3.57	3.29	77.36	3.30	3.24	58.70	3.34	3.32	54.86
98	3.76	3.59	67.84	3.81	3.59	72.12	3.55	3.58	50.30	3.71	3.72	52.86
99	4.27	4.06	68.98	4.34	4.07	72.94	4.01	4.14	42.62	4.24	4.40	42.78

Table 15. Q-Tests for Autocorrelations in Factor Returns

This table presents results of Ljung-Box Q-test up to 15 lags for returns on the four factors: *Market, Size (SMB), Value (HML),* and *Momentum*. In the table, *AC* refers to the autocorrelation, *Q* is the Q-test statistic and *Prob* represents the p-value of the test. We obtain the returns on the four factors (*Market, HML, SMB* and *Momentum*) over the period from January 1984 to September 2006 from French's online data library. We make the autocorrelation coefficients bold if they are no less than 10%.

	Market			SMB			HML			Momentum		
Lags	AC	Q	Prob	AC	Q	Prob	AC	Q	Prob	AC	Q	Prob
1	0.04	0.35	0.56	-0.03	0.29	0.59	**0.10**	2.62	0.11	-0.04	0.52	0.47
2	-0.06	1.22	0.54	0.02	0.44	0.80	0.05	3.19	0.20	-0.06	1.66	0.44
3	-0.05	1.99	0.58	**-0.13**	5.31	0.15	0.06	4.30	0.23	0.03	1.93	0.59
4	**-0.11**	5.47	0.24	-0.05	6.04	0.20	0.07	5.58	0.23	-0.09	4.14	0.39
5	0.04	5.92	0.31	0.05	6.80	0.24	-0.05	6.40	0.27	-0.02	4.29	0.51
6	0.04	6.31	0.39	0.01	6.84	0.34	0.04	6.93	0.33	**0.13**	9.37	0.15
7	0.04	6.80	0.45	0.08	8.60	0.28	0.04	7.36	0.39	-0.09	11.46	0.12
8	-0.05	7.37	0.50	-0.02	8.66	0.37	0.01	7.37	0.50	-0.04	11.91	0.16
9	-0.02	7.45	0.59	-0.02	8.81	0.46	-0.03	7.55	0.58	0.00	11.91	0.22
10	**0.11**	10.76	0.38	-0.03	9.14	0.52	0.05	8.33	0.60	-0.07	13.22	0.21
11	0.01	10.80	0.46	-0.04	9.56	0.57	0.09	10.43	0.49	**-0.10**	15.83	0.15
12	-0.03	11.01	0.53	0.01	9.62	0.65	-0.05	11.06	0.52	**0.19**	26.25	0.01
13	-0.01	11.06	0.61	0.03	9.90	0.70	-0.06	12.03	0.53	0.01	26.29	0.02
14	0.02	11.18	0.67	0.01	9.92	0.77	0.04	12.56	0.56	-0.04	26.64	0.02
15	-0.04	11.68	0.70	-0.04	10.46	0.79	-0.09	14.96	0.45	-0.04	27.02	0.03

Figure 1. PDFs of Alpha Estimates (DGP 1)

This figure plots the kernel density estimates of probability density functions (PDF) of realized (dashed line) and bootstrapped (solid line) alphas for DGP 1, in which, all funds are endowed with zero alphas. Details on how to obtain realized and bootstrapped alphas are given in Sections 2.2 and 2.3. In our simulation, fund returns are generated from the market model (1), in which, market returns are drawn randomly from $N(0.08, 0.15^2)$, βs are simulated from $U[0.5, 2]$ and errors are sampled randomly from one of the four distributions: uniform, normal, t and NIG. Parameters of the four distributions are set so that errors have zero mean and annual standard deviation 0.08. Specifically, the uniform distribution lies on $[-0.04, 0.04]$, the t distribution has a degree of freedom 3, and the NIG distribution has tail heaviness λ=10, location μ=0, asymmetry parameter θ=0 and scale parameter δ=0.0053. The four subplots are the estimated PDFs corresponding to the four different error distributions.

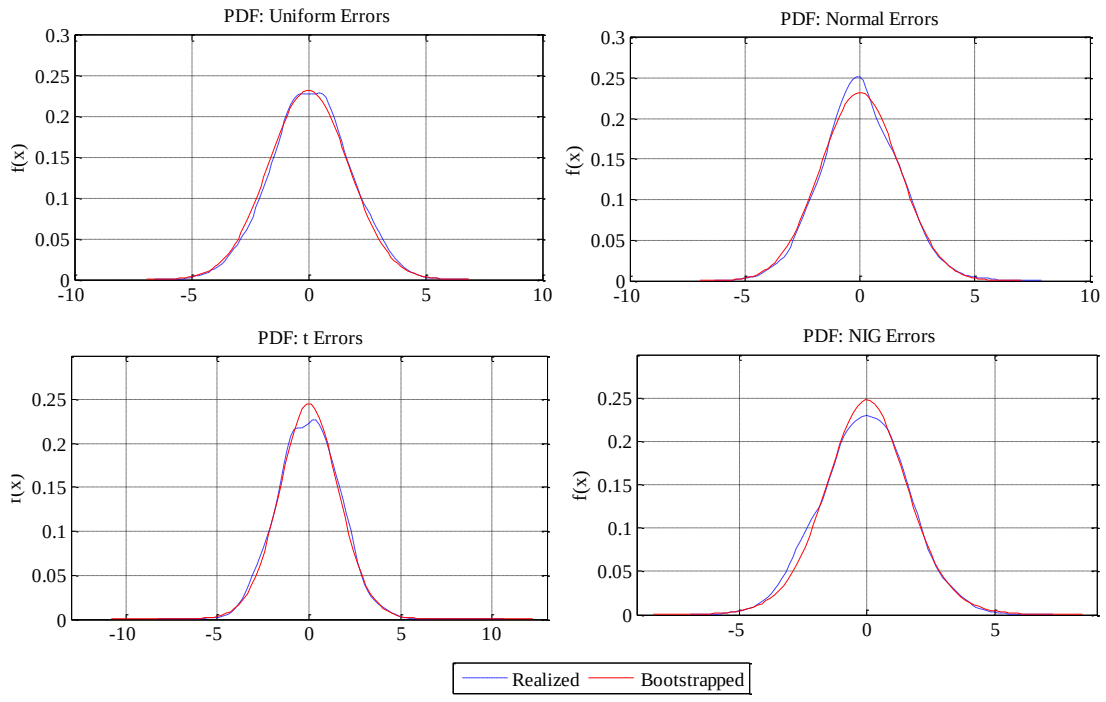


Figure 2. PDFs and CDFs of Alpha Estimates (the Four-Factor Model, DGP 1)

This figure plots kernel density estimates of cross-sectional distributions for the realized (dashed line) and bootstrapped (solid line) alphas for DGP 1: all funds are endowed with zero alphas. The eight subplots are estimated PDFs and CDFs under four different error distributions: uniform, normal, t and NIG, respectively. Details on how to obtain alpha estimates are given in Sections 2.2 and 2.3. We download factor returns over the period January 1984 to September 2006 from French's online data library. βs are sampled from the uniform distribution on [0.5, 2] and errors are drawn from one of the four distributions: uniform, normal, t and NIG. Parameters of the four distributions are set so that errors have zero mean and an annual standard deviation 0.08. Specifically, the uniform distribution lies on $[-0.04,0.04]$, the t distribution has a degree of freedom 3 and the NIG distribution has tail heaviness λ=10, location μ=0, asymmetry parameter θ=0 and scale parameter δ=0.0053.

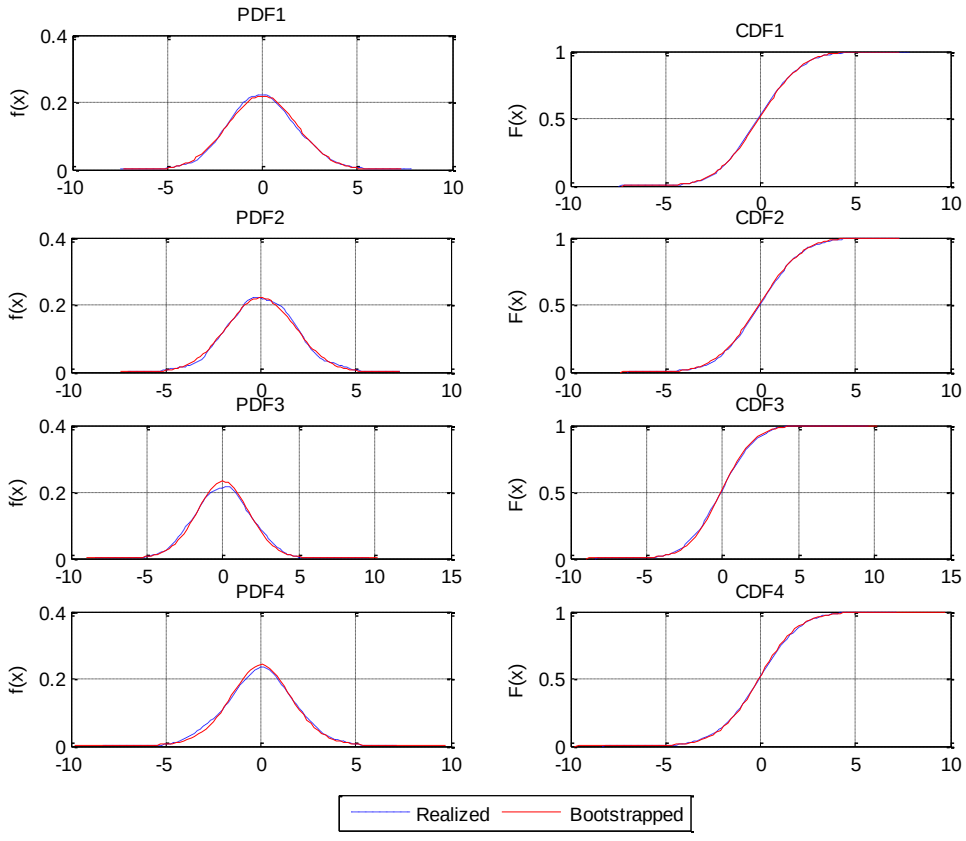
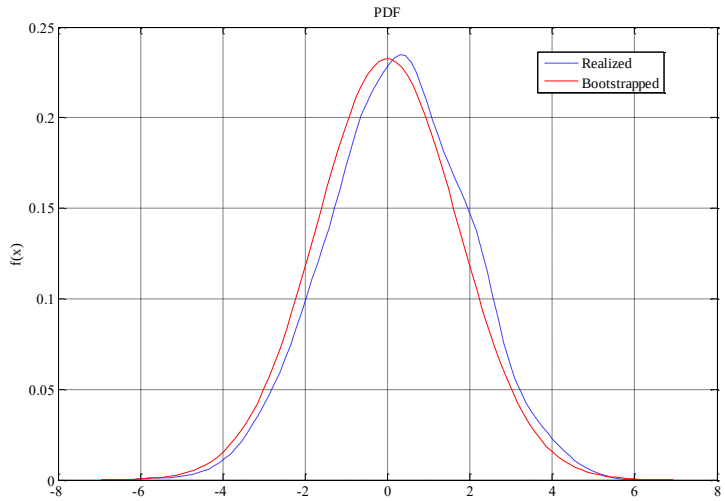


Figure 3. PDFs of Alpha Estimates in the Presence of Omitted Factor (DGP 1)

The figure shows kernel density estimates of probability density functions (PDFs) of realized (dashed line) and bootstrapped (solid line) alphas for DGP 1: all funds are endowed with zero alphas. Details on how to obtain percentiles for the t-statistics of alpha estimates are given in Sections 2.2 and 2.3. In our simulation, excess returns on the market factor $r_{m,t+1}$ are generated from (19), in which, γ is set to be 0.5, u_{t+1} is simulated from $N(0, 0.08^2)$ and Z_t is sampled from $N(0, 0.15^2)$. Fund returns are generated from the conditional beta model in (17), in which, factor loadings βs and errors $\varepsilon_{i,t}$ are drawn randomly from $U[0.5, 2]$ and $N(0, 0.08^2)$, respectively.



References

Abramowitz, M., 1974. Handbook of mathematical functions, with formulas, graphs, and mathematical tables (Dover Publications).

Barndorff-Nielsen, O. E., 1977, Exponentially decreasing distributions for the logarithm of particle size, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 353, 401-419.

Barndorff-Nielsen, O. E., 1997, Normal inverse Gaussian distributions and stochastic volatility modeling, Scandinavian Journal of Statistics 24, 1-13.

Barras, L., O., Scaillet, and R., Wermers, 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. Journal of Finance, 65, 179-216.

Berk, J. B., and J. H. van Binsbergen, 2015. Measuring skill in the mutual fund industry. Journal of Financial Economics, 118, 1-20.

Berk, J. B., and R. C., Green, 2004, Mutual fund flows and performance in rational markets. Journal of Political Economy, 112, 1269-1295.

Blake, D., A. G., Rossi, A., Timmermann, I.,Tonks, and R., Wermers, 2013, Decentralized investment management: Evidence from the pension fund industry. Journal of Finance, 68, 1133-1178.

Blake, D., Caulfield, T., Ioannidis, C., and I. P. Tonks, 2014. Improved inference in the evaluation of mutual fund performance using panel bootstrap methods. Journal of Econometrics, 183, 202-210.

Blake, D., Caulfield, T., Ioannidis, C., and I. P. Tonks, 2017. New evidence on mutual fund performance: A comparison of alternative bootstrap methods. Journal of Financial and Quantitative Analysis, 52, 1279-1299.

Carhart, M. M., 1997, On persistence in mutual fund performance, Journal of Finance 52, 57-82.

Chen, H. L., N. Jegadeesh, and R. Wermers, 2000, The value of active mutual fund management: An examination of the stockholdings and trades of fund managers, Journal of Financial and Quantitative Analysis 35, 343-368.

Chen, Y., Cliff, M., and H. Zhao, 2017. Hedge funds: The good, the bad, and the lucky. Journal of Financial and Quantitative Analysis, 52, 1081-1109.

Cheng, T., and C. Yan, 2017. Evaluating the size of the bootstrap method for fund performance evaluation. Economics Letters, 156, 36-41.

Christopherson, J., W. Ferson, and D. Glassman, 1998, Conditioning manager alphas on economic information: Another look at the persistence of performance, Review of Financial Studies 11, 111-142.

Cuthbertson, K., D. Nitzsche, and N. O'Sullivan, 2008, UK mutual fund performance: Skill or luck?, Journal of Empirical Finance 15, 613-634.

Eberlein, E., and U. Keller, 1995, Hyperbolic distributions in finance, Bernoulli 1, 281-299.

Elton, E. J., M. J. Gruber, S. Das, and M. Hlavka, 1993, Efficiency with costly information: A reinterpretation of evidence from managed factors, Review of Financial Studies 6, 1-22.

Fama, E. F., and K. R. French, 2010, Luck versus skills in the cross-section of mutual fund returns, Journal of Finance 65, 1915-1947.

Ferson, W. E. 2010. Investment performance evaluation. Annual Review of Financial Economics, 2, 207-234.

Ferson, W., and Y. Chen, 2017. How many good and bad fund managers are there, really? University of Southern California working paper.

Ferson, W. E., and R. W. Schadt, 1996, Measuring fund strategy and performance in changing economic conditions, Journal of Finance 51, 425-461.

Harvey, C. R., and Y. Liu, 2017. Rethinking performance evaluation. National Bureau of Economic Research working paper (No. w22134).

Henriksson, R. D., and R. C. Merton, 1981, On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills, Journal of Business 54, 513-533.

Jagannathan, R., and Z. Wang, 1996, The conditional CAPM and the cross-section of expected returns, Journal of Finance 51, 3-53.

Jensen, M. C., 1968, The performance of mutual funds in the period 1945-1964, Journal of Finance 23, 389-416.

Greene, W. H. 2011. Econometric analysis (7th), Prentice Hall, Englewood Cliffs, NJ.

Kosowski, R., A. Timmermann, R. Wermers, and H. White, 2006, Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis, Journal of Finance 61, 2551-2595.

Kosowski, R., N. Y., Naik, and M., Teo. 2007, Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. Journal of Financial Economics, 84, 229-264.

Malkiel, B. G., 1995, Returns from investing in equity mutual funds 1971 to 1992, Journal of Finance, 50, 549–572.

Newey, W. K., and K. D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, Econometrica 55, 703-708.

Politis, D. N., and J. P. Romano, 1994, The stationary bootstrap, Journal of the American Statistical Association 89, 1303-1313.

Treynor, J. L., and K. Mazuy, 1966, Can mutual funds outguess the market, Harvard Business Review 44, 131-136.

Wermers, R., 2000, Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses, Journal of Finance 55, 1655-1695.

Wermers, R. 2011. Performance measurement of mutual funds, hedge funds, and institutional accounts. Annual Review of Financial Economics, 3, 537-574.

Appendix: The Normal-Inverse Gaussian Distribution

Developed by Barndorff-Nielsen (1977), the Normal-Inverse Gaussian distribution (NIG) has been widely used in modeling financial asset returns, see Eberlein and Keller (1995) and Barndorff-Nielsen (1997). The density function for the NIG family can be expressed as

$$f(x;\lambda,\theta,\mu,\delta)=\frac{\lambda}{\pi\delta}\exp(\delta\sqrt{\lambda^2-\theta^2}+\theta(x-\mu))\frac{K_1(\lambda\delta\sqrt{1+\left(\frac{x-\mu}{\delta}\right)^2}}{\sqrt{1+\left(\frac{x-\mu}{\delta}\right)^2}}$$

where $x\in R,\lambda>0,\delta>0,\mu\in R,0<|\theta|<\lambda$ and K_1 is the modfied Bessel function of the third kind with index 1 (see Abramowitz (1974)).

The parameters μ, λ, θ and δ control respectively the location, tail heaviness, skewness and the scale of the distribution. The Gaussian distribution obtains as λ goes to infinity.

This class of distributions is characterized by the first four moments

$$E[x]=\mu+\frac{\delta\theta}{\sqrt{\lambda^2-\theta^2}}$$

$$Var[x]=\frac{\delta\lambda^2}{\left(\sqrt{\lambda^2-\theta^2}\right)^3}$$

$$Skew[x]=\frac{3\theta}{\lambda\sqrt{\delta\sqrt{\lambda^2-\theta^2}}}$$

$$Excess\ Kurtosis[x]=\frac{3}{\delta\sqrt{\lambda^2-\theta^2}}\left(1+\frac{4\theta^2}{\lambda^2}\right)$$

In our simulation, μ and θ are set to be zero and λ and δ are chosen to match the assumed variance.